Analysis Studio®

For Windows®

Statistical Analysis and Data Mining for Managers and Researchers

User Manual

Version 6

Appricon Inc.
Web Site: http://www.appricon.com
Email contact: info@appricon.com

Legal Statement
No part of this manual may be stored in a retrieval system, transmitted, or reproduced in any way, including but not limited to photocopy, photographs, magnetic, electronic or other method, without written permission from the publisher.
Appricon Inc. makes no guarantees with respect to the software and documentation and specifically disclaims any implied warranties of fitness and compatibility for any particular purpose.

Microsoft, Windows, Oracle, Excel are trademarks of their respective owners.

Analysis6 ® is a Registered Trademark.
Copyright © 2005-2010 Appricon Inc.

TABLE of CONTENTS

INTRODUCTION 1
ANALYSIS STUDIO FRAMEWORK 2
ANALYSIS STUDIO PROGRAM INSTALLATION 3
System Requirements 3.1
Analysis Studio Installation 3.2
Registration 3.3
Regional Settings Support 3.4
Treating Missing Values 3.5
Data Sets Features 3.6
ANALYSIS STUDIO PROGRAM OPERATION 4
Main Menu Bar 4.1
File Menu 4.1.1
New 4.1.1.1
Open 4.1.1.2
Save As 4.1.1.3
Save 4.1.1.4
Print 4.1.1.5
Export Tab 4.1.1.6
Exit 4.1.1.8
Edit Menu 4.1.2
Project Menu 4.1.3
Add Data Source 4.1.3.1
Delete Current Item 4.1.3.2
Data Menu 4.1.4
Add Variable 4.1.4.1
Variable Properties 4.1.4.2
Filter 4.1.4.3
Creating a Filter 4.1.4.4
Applying and Deleting a Filter from a Multiple Data Sources Project 4.1.4.5
Transform 4.1.4.6
Function Editor Screen 4.1.4.7
Multiple Variables Transformation 4.1.4.8
Dummy Variables 4.1.4.9
Statistics Menu 4.1.5
Brief Analysis 4.1.5.1
Frequency / Histogram 4.1.5.2
Data Correlation 4.1.5.3
Auto Data Correlation 4.1.5.4
Data Correlation and Auto Correlation Test Framework Display 4.1.5.5
Simple Regression (Explore Variable Relations) 4.1.5.6
Simple Regression Analysis 4.1.5.7
Multiple Regression (Explore Multiple Variable Relations) 4.1.5.9
Logistic Regression 4.1.5.10
Logistic Regression (Fractional Polynomials) 4.1.5.11
Cox Regression 4.1.5.12
Time Series and Forecasting 4.1.5.13
Cross-Tab tables 4.1.5.14
Charts Menu 4.1.6
New Chart 4.1.6.1
Legend Location and Format 4.1.6.2
Percentage Display 4.1.6.3
Tools 4.1.8
Help 4.1.9

1 INTRODUCTION

Analysis Studio is statistical software designed by Appricon Inc. to serve managers, executives and researchers from different fields of interest.

The software was designed to serve the needs of business professionals who seek scientific validation for their decision-making processes.

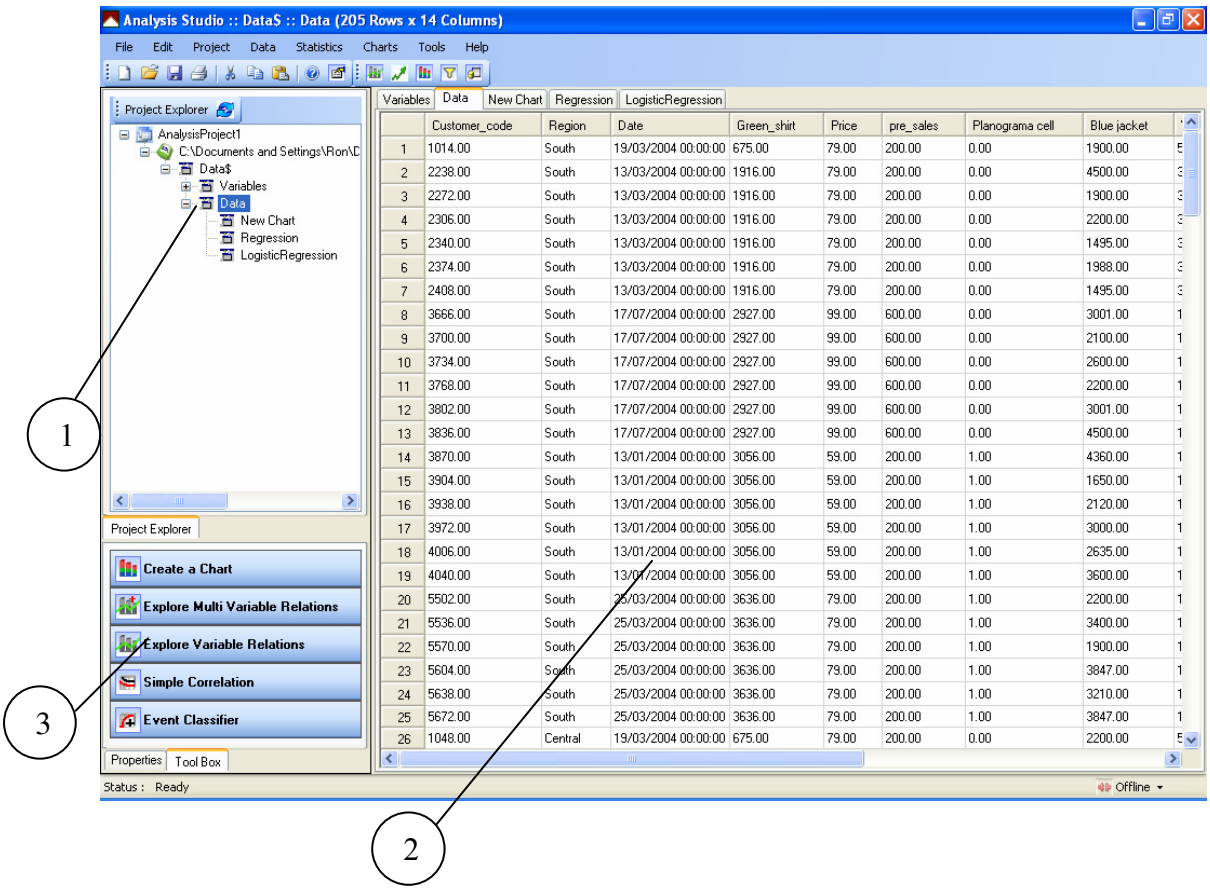
A unique and user-friendly interface also enables business professionals to explore and solve common challenges, quickly and accurately.

With Analysis Studio you can:

- ✓ Build a demand curve for your products
- ✓ Predict sales levels
- ✓ Identify customers with high churn likelihood
- ✓ Understand relationships between multiple-variables such as number of working hours, salary and professional rank
- ✓ Understand factors that lead to employee satisfaction and retention
- ✓ Reduce costs by analyzing hidden relationships between raw materials and production
- ✓ Predict profit levels based on accurate predictions of cost and sale revenues
- ✓ Identify profit failures before they occur

Analysis Studio also provides the tools for solving countless other optimization challenges in a quantitative (precise) manner, as opposed to a qualitative (assumption) manner.

2 ANALYSIS STUDIO FRAMEWORK



	Customer_code	Region	Date	Green_shirt	Price	pre_sales	Planograma cell	Blue jacket
1	1014.00	South	19/03/2004 00:00:00	675.00	79.00	200.00	0.00	1900.00
2	2238.00	South	13/03/2004 00:00:00	1916.00	79.00	200.00	0.00	4500.00
3	2272.00	South	13/03/2004 00:00:00	1916.00	79.00	200.00	0.00	1900.00
4	2306.00	South	13/03/2004 00:00:00	1916.00	79.00	200.00	0.00	2200.00
5	2340.00	South	13/03/2004 00:00:00	1916.00	79.00	200.00	0.00	1495.00
6	2374.00	South	13/03/2004 00:00:00	1916.00	79.00	200.00	0.00	1988.00
7	2408.00	South	13/03/2004 00:00:00	1916.00	79.00	200.00	0.00	1495.00
8	3666.00	South	17/07/2004 00:00:00	2927.00	99.00	600.00	0.00	3001.00
9	3700.00	South	17/07/2004 00:00:00	2927.00	99.00	600.00	0.00	2100.00
10	3734.00	South	17/07/2004 00:00:00	2927.00	99.00	600.00	0.00	2600.00
11	3768.00	South	17/07/2004 00:00:00	2927.00	99.00	600.00	0.00	2200.00
12	3802.00	South	17/07/2004 00:00:00	2927.00	99.00	600.00	0.00	3001.00
13	3836.00	South	17/07/2004 00:00:00	2927.00	99.00	600.00	0.00	4500.00
14	3870.00	South	13/01/2004 00:00:00	3056.00	59.00	200.00	1.00	4360.00
15	3904.00	South	13/01/2004 00:00:00	3056.00	59.00	200.00	1.00	1650.00
16	3938.00	South	13/01/2004 00:00:00	3056.00	59.00	200.00	1.00	2120.00
17	3972.00	South	13/01/2004 00:00:00	3056.00	59.00	200.00	1.00	3000.00
18	4006.00	South	13/01/2004 00:00:00	3056.00	59.00	200.00	1.00	2635.00
19	4040.00	South	13/01/2004 00:00:00	3056.00	59.00	200.00	1.00	3600.00
20	5502.00	South	25/03/2004 00:00:00	3636.00	79.00	200.00	1.00	2200.00
21	5536.00	South	25/03/2004 00:00:00	3636.00	79.00	200.00	1.00	3400.00
22	5570.00	South	25/03/2004 00:00:00	3636.00	79.00	200.00	1.00	1900.00
23	5604.00	South	25/03/2004 00:00:00	3636.00	79.00	200.00	1.00	3847.00
24	5638.00	South	25/03/2004 00:00:00	3636.00	79.00	200.00	1.00	3210.00
25	5672.00	South	25/03/2004 00:00:00	3636.00	79.00	200.00	1.00	3847.00
26	1048.00	Central	19/03/2004 00:00:00	675.00	79.00	200.00	0.00	2200.00

The Analysis Studio screen has three main frames:

Frame 1: Project Explorer – Includes a project framework, its data source paths, current statistical tabs and names and filters of project variables.

Frame 2: Operational Display – Includes the exchangeable tabs at the top. By clicking on a tab, its linked data will appear.

Frame 3: Tool Box – Contains the quick launch exchangeable option for the main statistical procedures.

3 ANALYSIS STUDIO PROGRAM INSTALLATION

3.1 System Requirements

To run Analysis Studio, you need an IBM-compatible computer with a Pentium 3 or an equivalent processor or better, at least 128 MB of memory, a mouse, Windows 2000, XP or later, .NET Framework component, and 100 Megabyte free space on your hard disk. Note: please use Windows Update® to obtain .NET Framework or click the link:

3.2 Analysis Studio Installation

Note: To install Analysis Studio on Windows 2000 or later, you must be logged in to your computer with administrator privileges.

1. After downloading the software package from the Appricon website, click the "Setup" button and follow the Wizard instructions.

2. When installation is complete, start Analysis Studio by clicking the "Start" button, pointing to Programs and selecting the Analysis Studio icon.

3.3 Registration

1. Enter your user name and product key.
   If you do not have a product key, you can purchase one from the Analysis Studio web site (http://www.appricon.com). If you are a registered user, and you have lost your product key you can contact Appricon (info@appricon.com), and we will email you your user name and product key.

2. In the first Analysis Studio dialog box, please enter the product key that was sent to you as part of the downloading process. You only have to enter your user name and product key once, the next time you start Analysis Studio the program will not ask you for this information again.

3.4 Regional Settings Support

The program supports regional differences as displayed in the Regional settings dialog box in Microsoft windows 2000/XP/VISTA/2003 server® Control frames.
The program supports the following formats from the Microsoft Windows settings:
Formatting symbols: All Windows characters are supported by the program.
Date formats: MM.DD.YY, DD.MM.YY or YY.MM.DD
Culture settings: The program supports over 170 languages and local currencies.

3.5 Treating Missing Values

The program calculates variables that have numeric values and ignores variables that contain text or date values. If there is a text value, the program ignores it and no calculation will be computed. The program supports legal separators (for example: 14,000.0, $14.00, -14.00) and ignores illegal separators (for example: 14'5, 14-5, "14")
At any case, the program will not transform missing values into zero. The program worksheet will display a missing value as N/A.

3.6 Data Sets Features

The program allows unlimited number of columns and rows. Please take into account that as more values are added to a data set more computation time will be necessary for statistical procedures. The program will display the data set and its attributes as the original data set file.

It is advised to have a formatting procedure made at the original data set file.

4 ANALYSIS STUDIO PROGRAM OPERATION

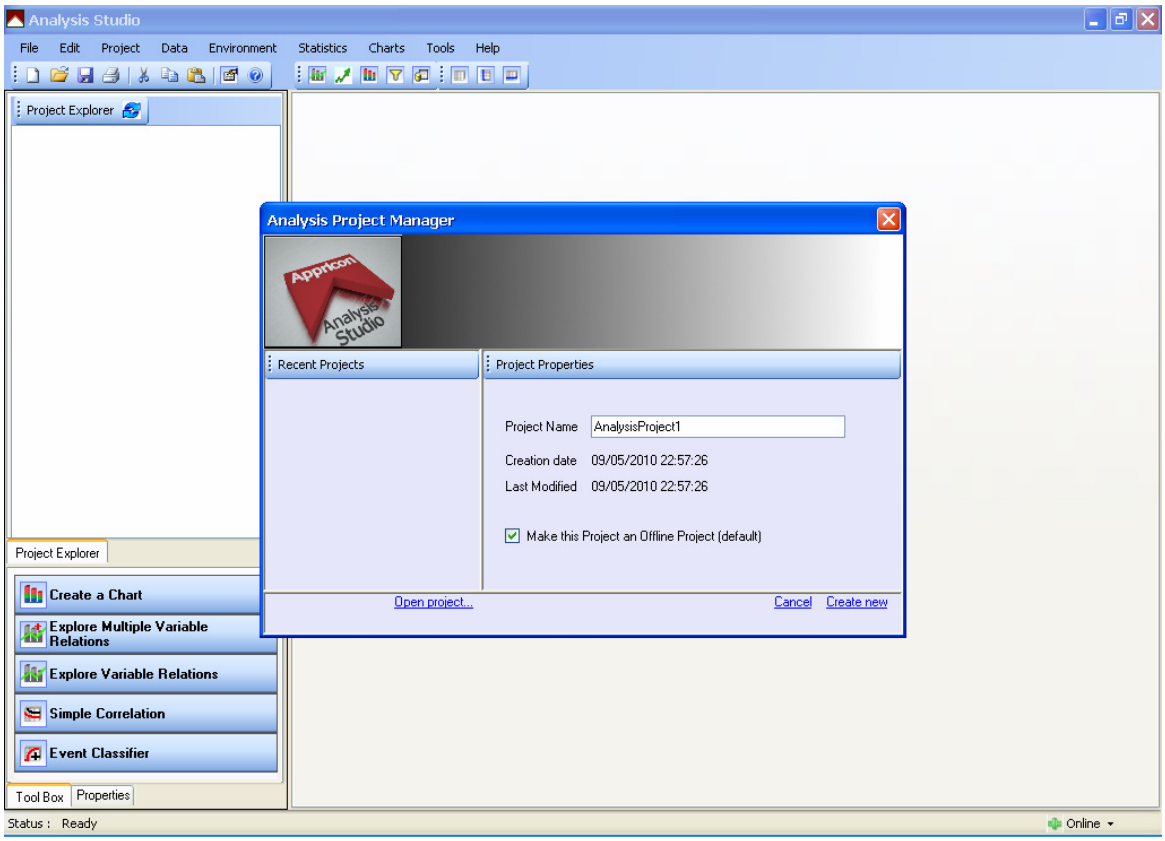
4.1 Main Menu Bar

The Main menu bar choices provide the following options:
File: New, Open, Save, Print and Exit
Edit: Data manipulations such as Copy, Cut, and Paste
Project: Add data source, Delete current item and Project properties
Data: Datasheet and columns properties.
Statistics: Statistical procedures such as regressions, significant tests, classification etc.
Charts: Chart creator
Tools: Layout options
Help: PDF help file and version info
Analysis Studio – User manual by Appricon 9 from 92

4.1.1 File menu

4.1.1.1 New

By clicking the File | New option, a dialog box with a default project name appears. A user may change it. For the stand-alone version of the software, the off-line check-box should stay on.

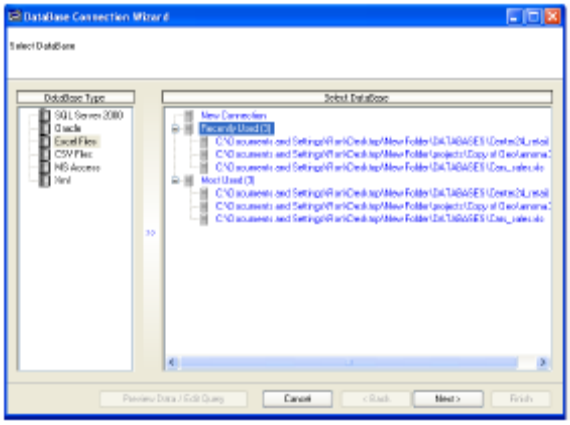


After clicking "OK", the database connection Wizard appears in order to provide an interface for pooling data from a file with one of the following known file extensions of

MSSQL 2000® or higher, Oracle 8.0® or higher, Excel 4.0® or higher, CSV, MSAccess 97® or higher and XML.

A user who wishes to pool data from SAS® or SPSS® or other file formats should first save the data in the XML/TEXT/CSV formats and then to use Analysis Studio® database connection Wizard for pooling the data.

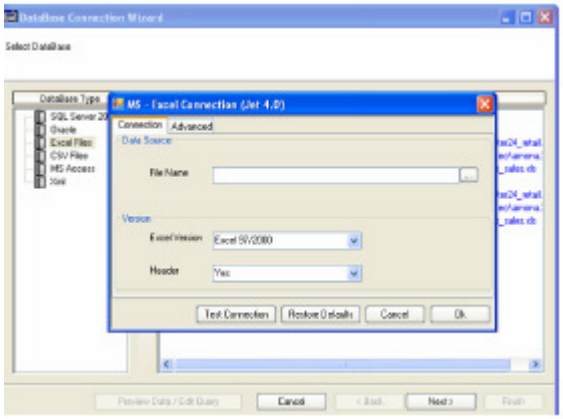
Database Connection 1: Extracting Data Out of an Excel 2000® File



After selecting an Excel file from “Database Type” at the left side of the screen, the three options appear on the right side of the screen:

→ New Connection: use this option to set a new connection to a data file.
→ Recently Used: displays the last eight files used, sorted by most recent use.
→ Most Used: displays the last eight files used, sorted by popularity.

Database Connection 2: Using a New Connection

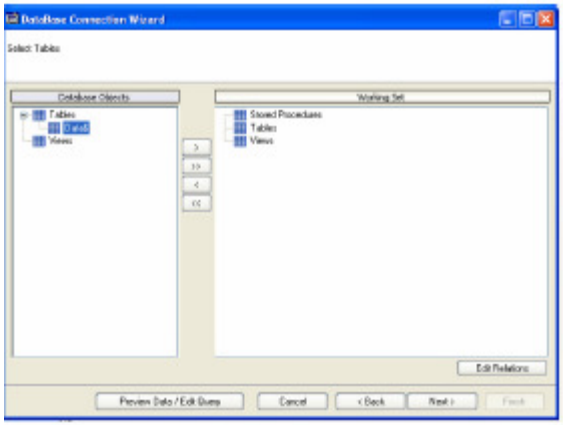


Use the File Name browse button to locate and select a desired file. Make sure that the file you select is the right Excel® version. Set the Header option to either “Yes” or “No”:

- If the Header option is set to "Yes", first row values of the file will turn into Analysis Studio variable names, one for each column.

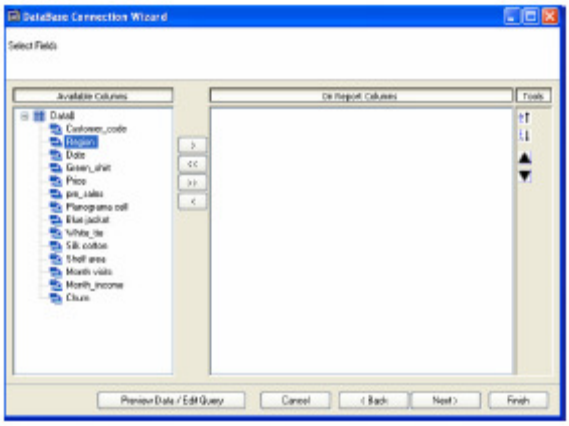
- If the Header option is set to "No", Analysis Studio will use its default names for each column.

Database Connection 3: Selecting a Spreadsheet



On the left side of the Wizard screen, a user chooses the desired spreadsheet (in the case of Excel ® or other single table or view file). A user can select only one table to be extracted. After selecting a desired spreadsheet, a user should use the selection arrows in order to move this spreadsheet to the Working Set frame that represents actual data for a display on the application grid.

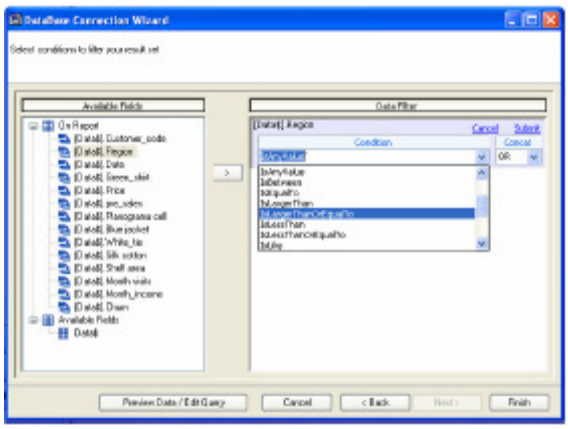
Database Connection 4: Selecting Columns from a Spreadsheet



It is possible to select all columns of a spreadsheet by double clicking the spreadsheet name (in this case "Data$") or to select one or more of its columns. After selecting desired columns, a user should use the selection arrows in order to move columns to the On Report Columns frame that represents actual columns for display on the application Data grid tab page.

A user can change the original order of columns (the first column of the list is the left column on the grid) or sort them by using the Tools frame features on the right side of the Wizard screen.

Database Connection 5: Filtering Selected Columns Data



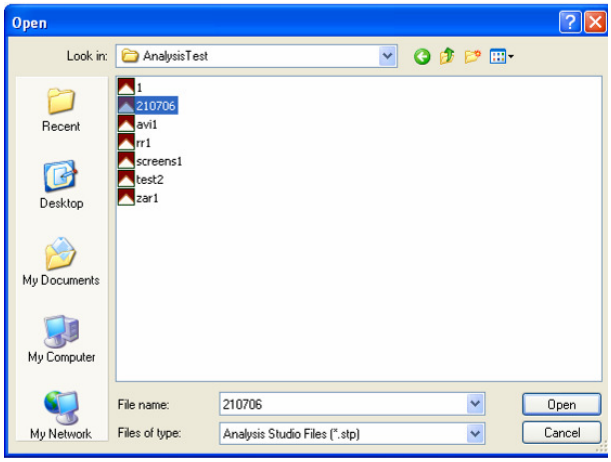
The user can use the Data Filter frame for excluding unnecessary data from a selected column (field). The first step is to select a desired column on the left frame (in this example: the Region column) by using the arrow button. The desired column header appears at the right frame of the Wizard. The second step is to compose its filter out of the proposed conditions. A multiple-layered filter is composed by using the "OR" and "And" operators. The Filter option is also available from the Main menu bar: Data | Filter.

4.1.1.2 Open

The “File Open” option serves for opening an Analysis Studio file named (*.stp) that represents a previously built project. The Wizard uses one dialog box for this purpose.

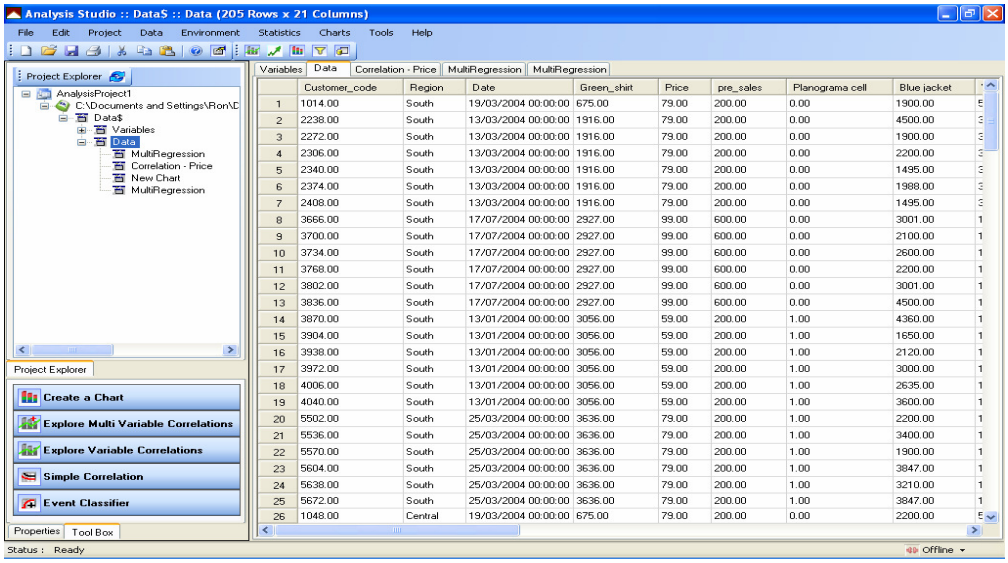
Open *stp file: Selecting the Desired Project

By selecting a *.stp file (In this example 210706.stp) the project data and all statistical procedures will be retrieved and displayed on the application screen.



Displaying the *stp file

The file is opened with all saved data and statistical procedures. The user can delete or include additional statistical procedures or any data manipulation and save everything to the current file or a new one.



4.1.1.3 Save As

The “Save As” option serves for saving data and statistical procedures of the current project. The project will be saved using two files: a *.data file that contains all data that were available at the time of saving and *.stp file that contains all statistical procedures that were available at the time of saving. If this is the first time a project is about to be saved the Save As dialog box will allow a user to write a desired name for a project and to select a folder that will contain the two project's files.

4.1.1.4 Save

This option allows for saving a project that has already been saved at least once during the current session. There is no dialog box for this purpose since the project is already saved according to the Save As settings.

4.1.1.5 Print

This option serves to print current statistical or chart tab data. Clicking on this option generates a report viewer that displays current selected tab data in a print preview mode.

The report viewer has its own menu bar that includes:

Pages Navigator, Refresh, Print Setup, Page Setup, Export to PDF or Excel and Zoom.

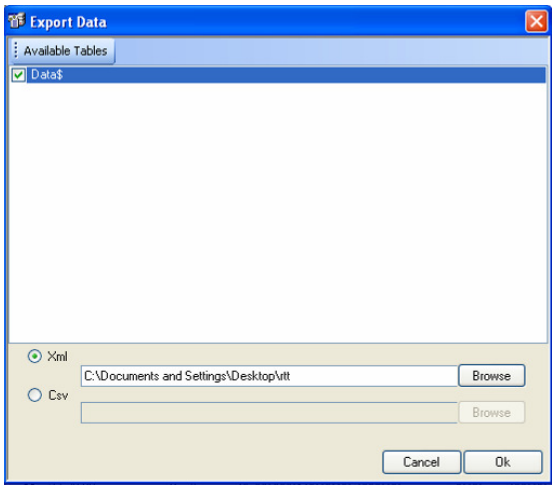
The picture above shows the Report Viewer menu bar.

4.1.1.6 Export Tab

There are two export options: “Export to Excel” and “Export to PDF” each of these options is designed for transformation of statistical or chart tabs into Excel or PDF format. To use these options the user should click on the desired tab and then to go to: File | Export Tab. In the Export Tab submenu, the user can select the desired export format.

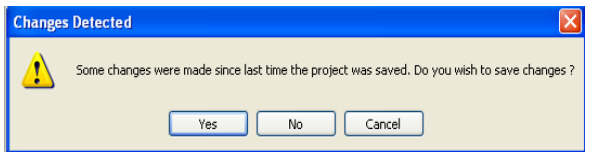
4.1.1.7 Export Data

This option exports the manipulated data (i.e. data that have Dummy variables or filters etc.) to a XML or CSV format for further use by other software tools.



4.1.1.8 Exit

The “Exit” procedure closes the entire software session. Prior to closing, a dialog box appears in order to assist the user in choosing the appropriate closing option.



4.1.2 Edit Menu

Editing data or variables (columns) includes deleting, adding, cutting, copying and pasting. The editing operations can be done before any statistical procedures are done. After performing a statistical operation the editing options will be disabled.

Cut

By clicking the “Cut” submenu option, a user can cut marked cells and paste their values into targeted cells on the Data tab grid or into Microsoft Excel® grid. If targeted cells on the Data grid have another format than the cells that were cut, a warning message will appear and the operation will be cancelled.

Copy

By clicking the “Copy” submenu option, a user can copy marked cells and paste their values into targeted cells on the Data grid or into Microsoft Excel® grid under the same constraint as mentioned above.

Paste

By clicking the “Paste” submenu option, copied or cut cells will be pasted into target cells under the same constrain as mentioned in the Cut frame.

4.1.3 Project Menu

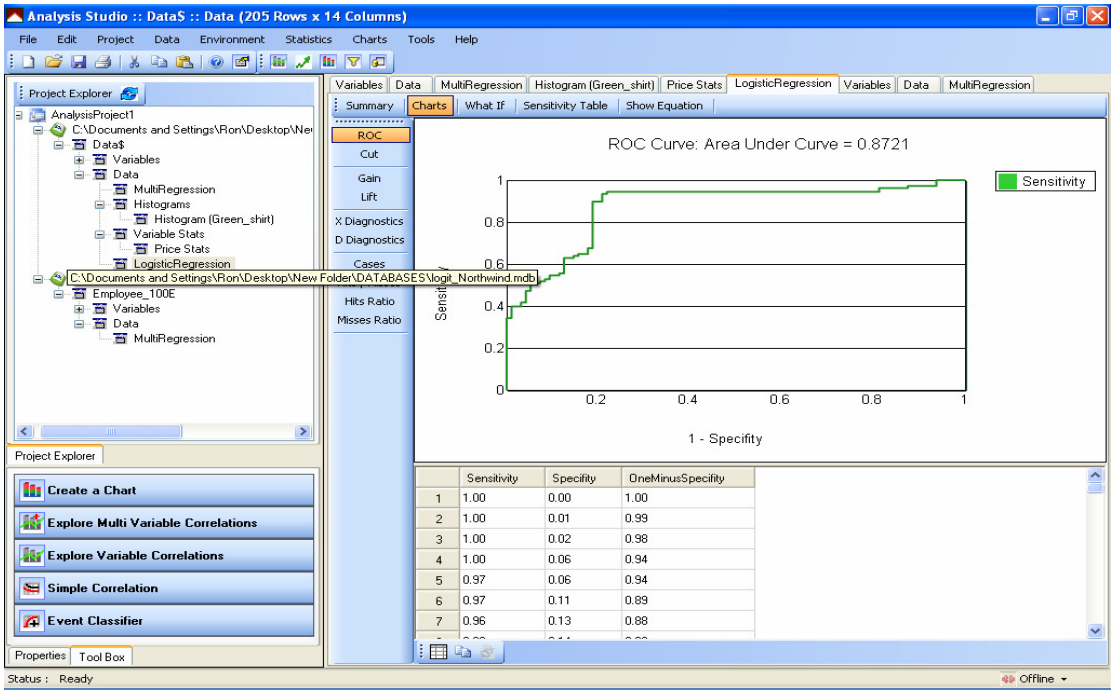
4.1.3.1 Add Data Source

Analysis Studio supports multiple-data sources for a project. The user can add as many Data Sources as needed. Each Data Source is represented with its own Variables and Data tab pages. A user can swiftly switch between two or more Data Sources and perform statistical inquiries for each Data Source within the same project.

To add a new Data Source to a current project:

In the Project submenu, select “Add Data Source”.

Working with Multiple Data Sources

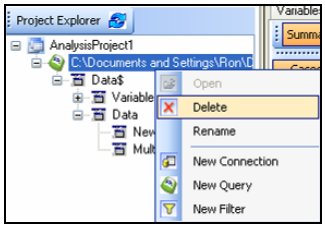


In the screenshot above, there are two Data Sources. The first one has four statistical procedures. The Logistic Regression is the current tab frame which is also marked in the

Project Explorer frame. The second Data Source has one statistical procedure which is multiple regression.

4.1.3.2 Delete Current Item

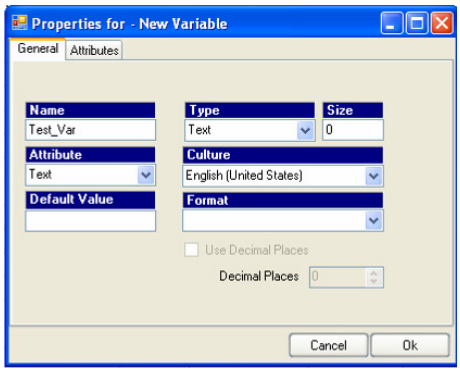
A user can delete one or more Data Source items. There are two ways for deleting a Data Source item. A user can right click a Data Source object in the Project Explorer frame and then select the Delete option. Alternatively, a user can select a Data Source object in the Project Explorer tree frame and then select the Delete submenu option from the Project submenu.



4.1.4 Data menu

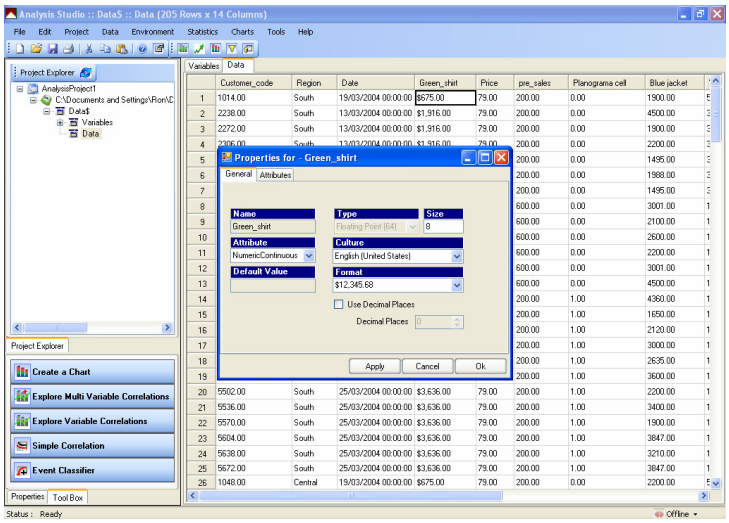
4.1.4.1 Add Variable

This option serves for adding a new variable to an existing data set. When clicking the *Add Variable* option, the *New Variable* dialog box opens and a user can set new variable settings such as Name, Type, Size, Culture, Format, Default value and Decimal places if the variable is a numeric type.



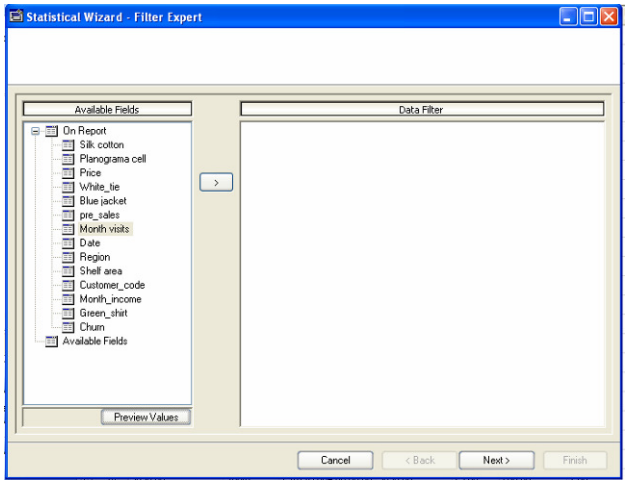
4.1.4.2 Variable Properties

This option is available for existing variables and is used to display variable properties while working with the Data tab window. A user can change variable properties as needed except for a variable Name and Type.



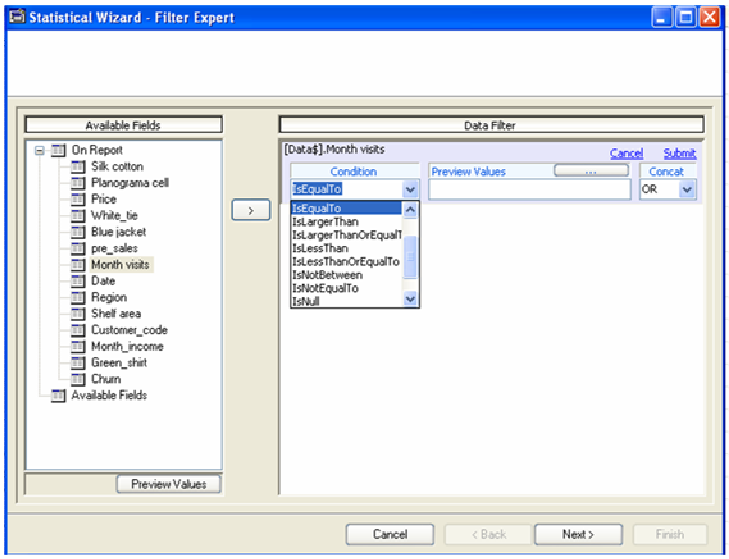
4.1.4.3 Filter

The *Filter* option serves for filtering existing data set rows (cases). By applying the Filter option, a user can change the data set as desired. By clicking the Filter sub-menu, the Filter expert Wizard appears:
A user can select the desired field (column) for filtering from the *Available Fields* on the left side of the window. The "Preview Values" button displays values for the selected field. By clicking on the arrow button in the middle, the selected field moves to the right side of the window and the filtering procedure can be created.

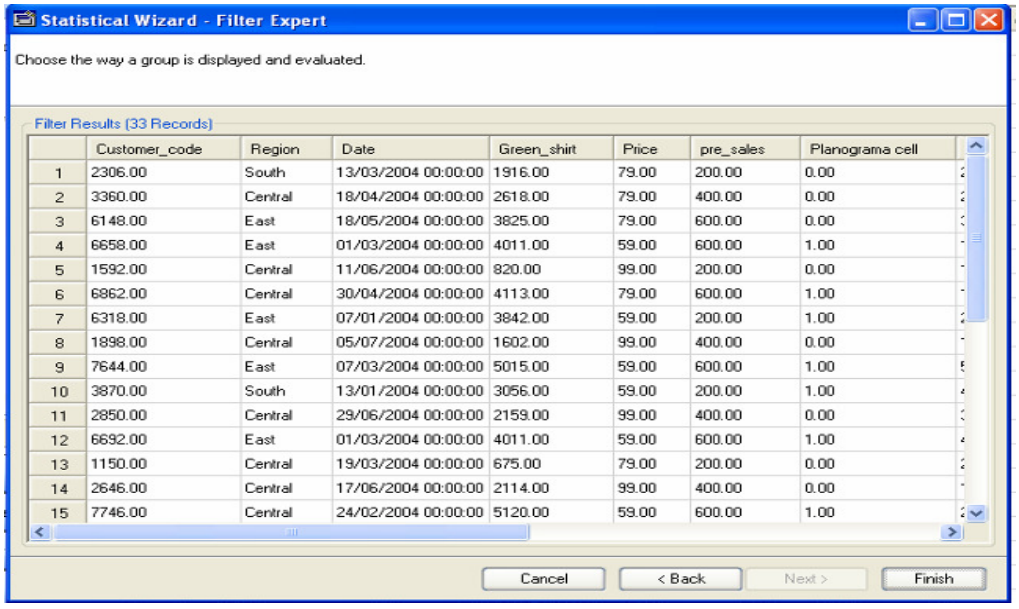


4.1.4.4 Creating a Filter

The Data Filter frame presents a filter that a user creates. A filter can have one condition or multiple conditions. A field name is displayed above the filter boxes.

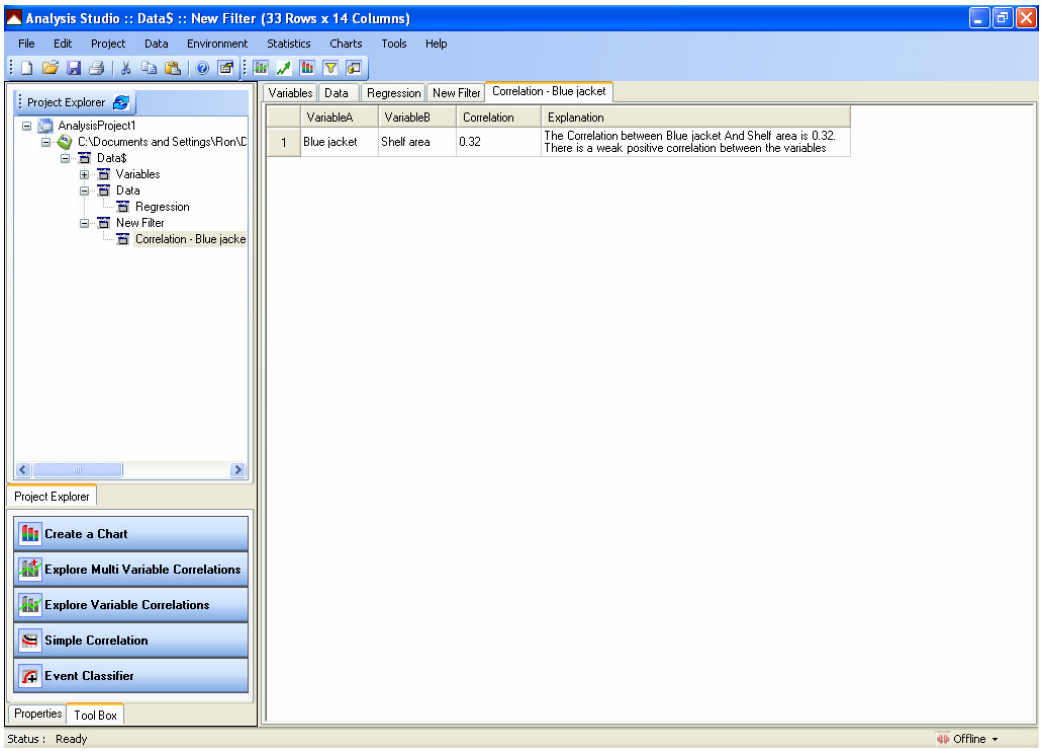


The *Condition* combo box stores 15 conditions that are available to choose from. The *Preview Values* text box is for selecting or writing values for the condition settings. The *Concat* (Concatenation) combo box serves for creating a multiple conditions filter. The use of "OR" in the Concat field is to separate conditions for the same filter and the use of "AND" is to combine one or more conditions with the first one. A filter condition can be removed, viewed or edited in order to change former condition properties. After creating a filter, the *filter results* screen appears and displays cases that are included in a new filtered data set.


Statistical Wizard - Filter Expert
Choose the way a group is displayed and evaluated.
Filter Results (33 Records)
Cancel
< Back
Next >
Finish

	Customer_code	Region	Date	Green_shirt	Price	pre_sales	Planograma cell
1	2306.00	South	13/03/2004 00:00:00	1916.00	79.00	200.00	0.00
2	3360.00	Central	18/04/2004 00:00:00	2618.00	79.00	400.00	0.00
3	6148.00	East	18/05/2004 00:00:00	3825.00	79.00	600.00	0.00
4	6658.00	East	01/03/2004 00:00:00	4011.00	59.00	600.00	1.00
5	1592.00	Central	11/06/2004 00:00:00	820.00	99.00	200.00	0.00
6	6862.00	Central	30/04/2004 00:00:00	4113.00	79.00	600.00	1.00
7	6318.00	East	07/01/2004 00:00:00	3842.00	59.00	200.00	1.00
8	1898.00	Central	05/07/2004 00:00:00	1602.00	99.00	400.00	0.00
9	7644.00	East	07/03/2004 00:00:00	5015.00	59.00	600.00	1.00
10	3870.00	South	13/01/2004 00:00:00	3056.00	59.00	200.00	1.00
11	2850.00	Central	29/06/2004 00:00:00	2159.00	99.00	400.00	0.00
12	6692.00	East	01/03/2004 00:00:00	4011.00	59.00	600.00	1.00
13	1150.00	Central	19/03/2004 00:00:00	675.00	79.00	200.00	0.00
14	2646.00	Central	17/06/2004 00:00:00	2114.00	99.00	400.00	0.00
15	7746.00	Central	24/02/2004 00:00:00	5120.00	59.00	600.00	1.00

A filtered data set is considered a subset of the main project data and has its own node on the Project Explorer left frame. Any new statistical procedure that performs on the filtered dataset will be displayed under the *New Filter* node.


Analysis Studio :: Data$ :: New Filter (33 Rows x 14 Columns)
File Edit Project Data Environment Statistics Charts Tools Help
Project Explorer
AnalysisProject1
C:\Documents and Settings\Ron\D
Data$
Variables
Data
Regression
New Filter
Correlation - Blue jacke
Variables Data Regression New Filter Correlation - Blue jacket
VariableA VariableB Correlation Explanation
1 Blue jacket Shelf area 0.32 The Correlation between Blue jacket And Shelf area is 0.32. There is a weak positive correlation between the variables
Project Explorer
Create a Chart
Explore Multi Variable Correlations
Explore Variable Correlations
Simple Correlation
Event Classifier
Properties Tool Box
Status : Ready
Offline

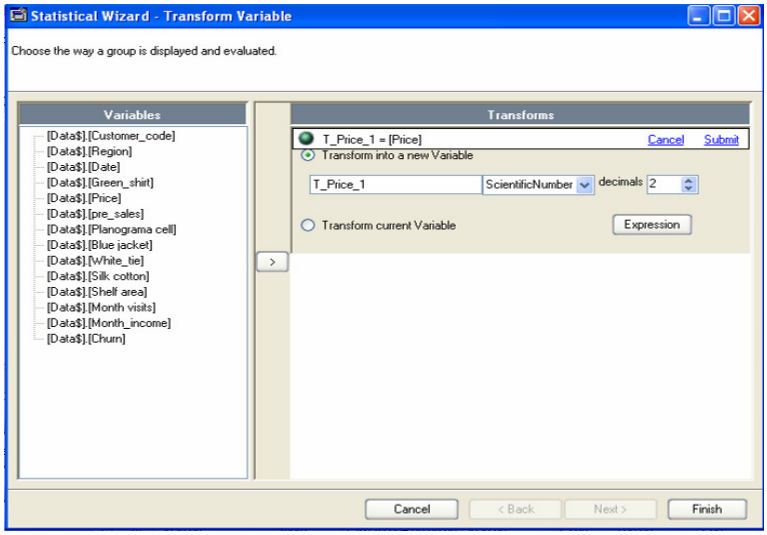
4.1.4.5 Applying and Deleting a Filter from a Multiple Data Sources Project

To create a filter from a multiple data sources project, a user should select the appropriate data node and select “Filter” from the Data menu or use the button. The procedure for deleting a filter from a multiple data sources project is the same as for a single data source.

4.1.4.6 Transform

Clicking the “Transform” option generates the variable transformation Wizard that assists in producing a new variable with the same properties as the original one, but with another name or, based on the original name but with some properties changed.

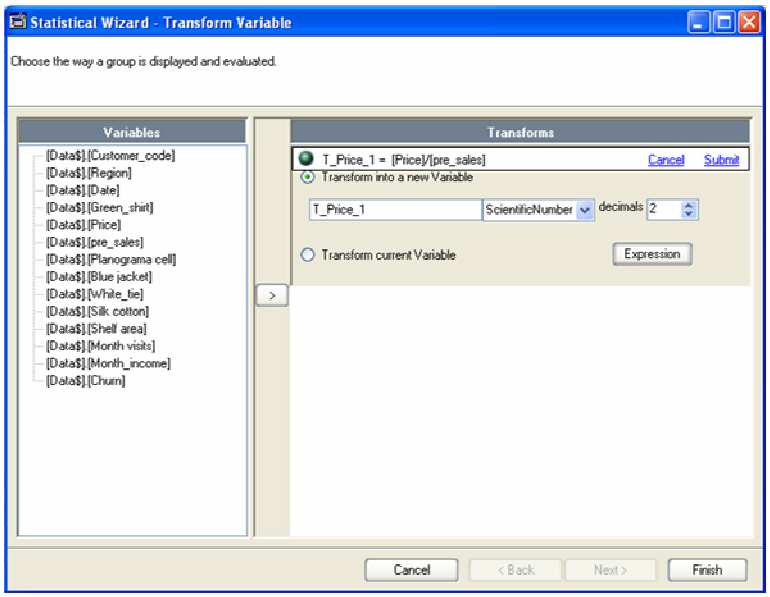
Transform Wizard



By selecting a desired variable on the left side of the window and clicking the arrow button, two options for transforming the selected variable are proposed.

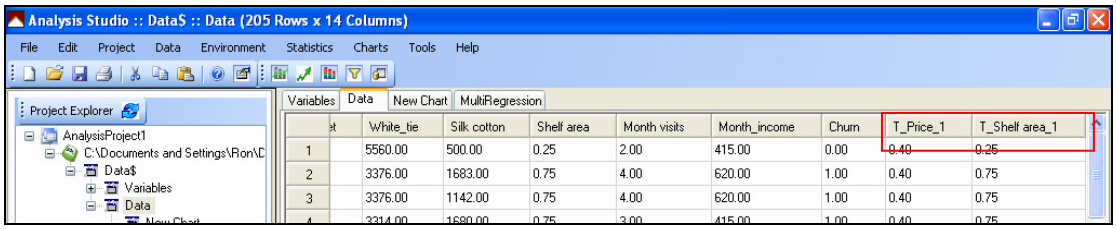
Transform into a new Variable: A user can create a new variable based on current variable properties. At any time, a user can change a default variable name, data type and desired decimal places.
Transform current Variable: A user can change variable properties by using the "Expression" button.

The Transform Wizard can deal with a multiple variables transformation by clicking the *Submit* button for each variable transformation.



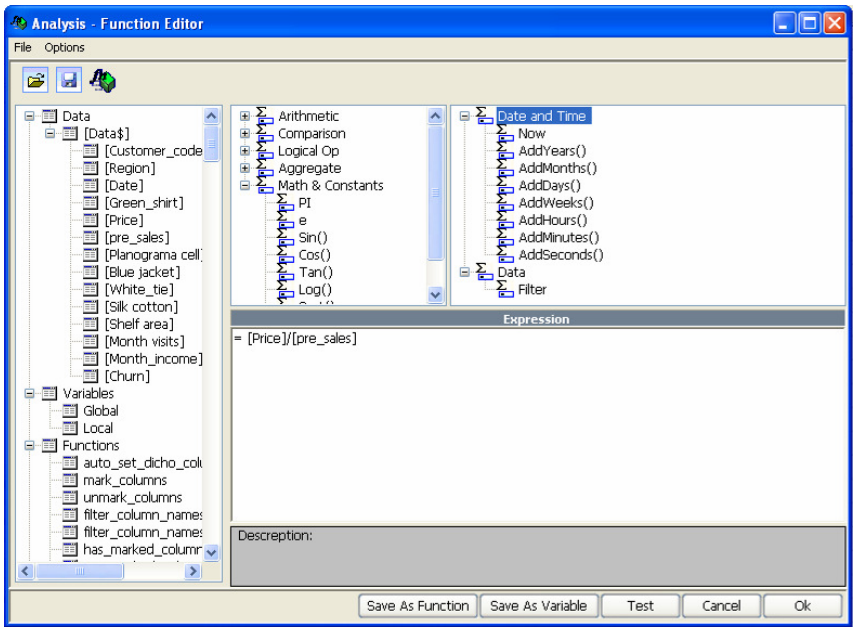
4.1.4.7 Multiple Variables Transformation

As the user clicks the “Finish” button, the new variable/variables are computed and presented on the Data tab grid as seen in the screenshot below.


Analysis Studio :: Data$ :: Data (205 Rows x 14 Columns)
File Edit Project Data Environment Statistics Charts Tools Help
Project Explorer
AnalysisProject1
Data$
Variables
Data
Variables Data New Chart MultiRegression
White_tie Silk cotton Shelf area Month visits Month_income Churn T_Price_1 T_Shelf area_1
1 5560.00 500.00 0.25 2.00 415.00 0.00 0.40 0.25
2 3376.00 1683.00 0.75 4.00 620.00 1.00 0.40 0.75
3 3376.00 1142.00 0.75 4.00 620.00 1.00 0.40 0.75

4.1.4.8 Function Editor Screen

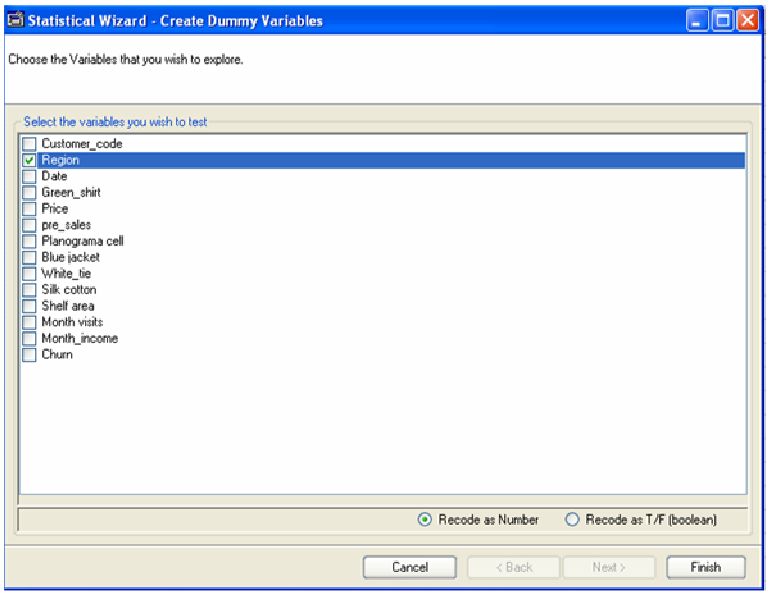
The function editor is a powerful creator of scripts and functions that can assign functional and logical expressions to variable values. In the screenshot above, a simple function is assigned to the “T_price_1 new variable”. By clicking the "OK" button, the new function becomes the default function of the new variable as seen in the screenshot below.



4.1.4.9 Dummy Variables

Dummy variables are also known as design variables and are used to transform non-numeric variables to numeric variables that can be included in statistical procedures like regressions of all types. For example: If one would like to understand the contribution of the "Region" variable on sales, transforming the "Region" non-numeric values (such as "East"; "North" etc.) into some numeric values (i.e. "East" will be transformed into "1" and "North" into "0") is a must.

Creating Dummy Variables: Step 1



Select the option “Dummy Variables” from the Data menu. A user should select a non-numeric variable/variables for a transformation. The Analysis Studio software automatically creates a dummy variable from an original one. Each value name of a dummy variable always has the same pattern: DUMMY_original variable name_original value. For example: from the "Region" variable Analysis Studio has created Dummy variable values that had four different values.

DUMMY_Region_null	DUMMY_Region_Central	DUMMY_Region_East	DUMMY_Region_South
0	0	0	1
0	0	0	1
0	0	0	1

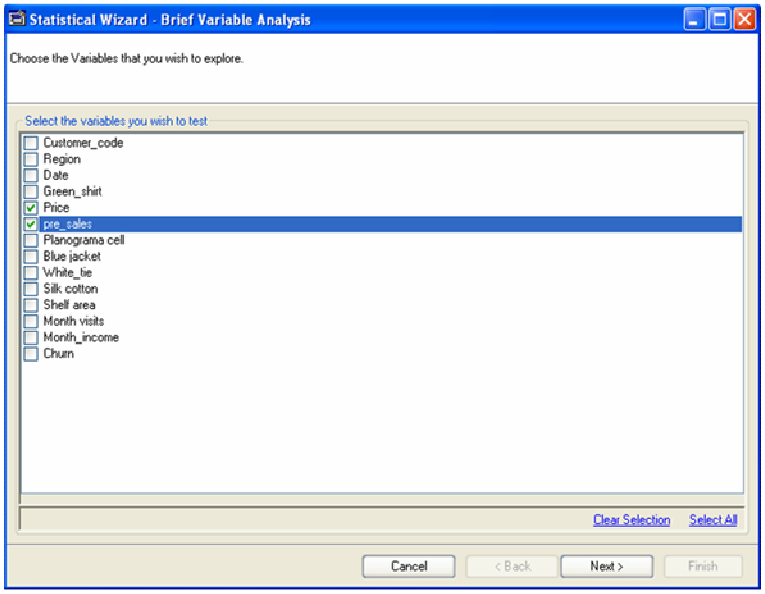
4.1.5 Statistics Menu

The Statistics menu contains the statistical procedures that Analysis Studio computes. The statistical Wizards and outputs are designed for researchers as well as business users. Each one of the statistical procedures has its own tab page and a unique node in the Project Explorer that makes navigation simple and clear yet very powerful for data manipulations.

4.1.5.1 Brief Analysis

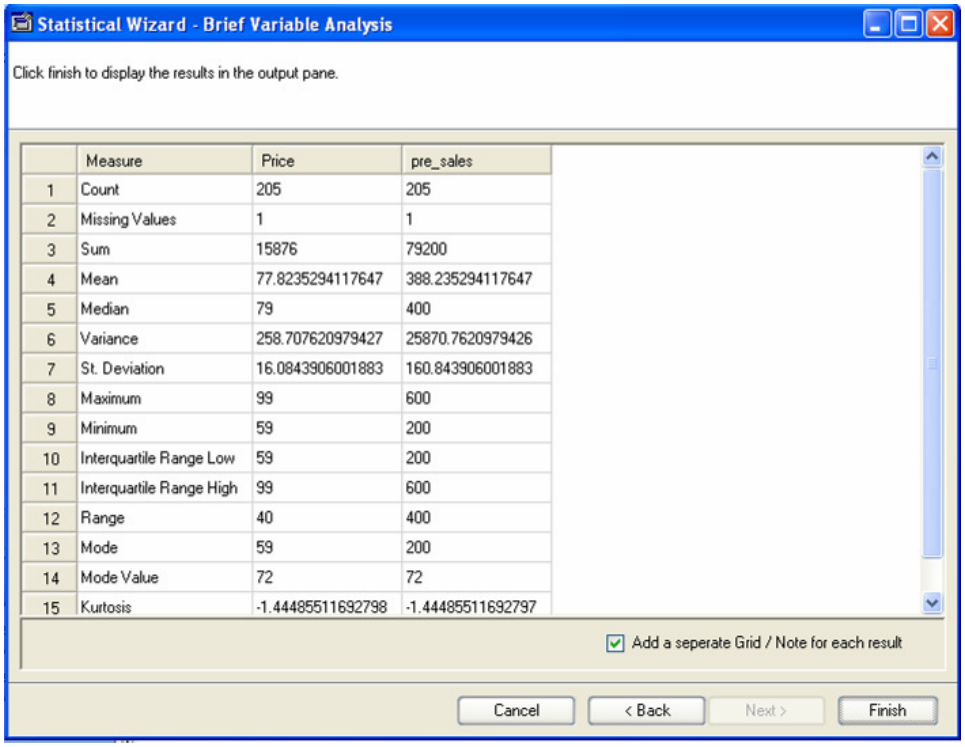
To get a clear insight into variable values the Brief Analysis procedure contains desired variable parameters including advanced statistical parameters.

Brief Variable Analysis Wizard Step 1: Selecting a Variable



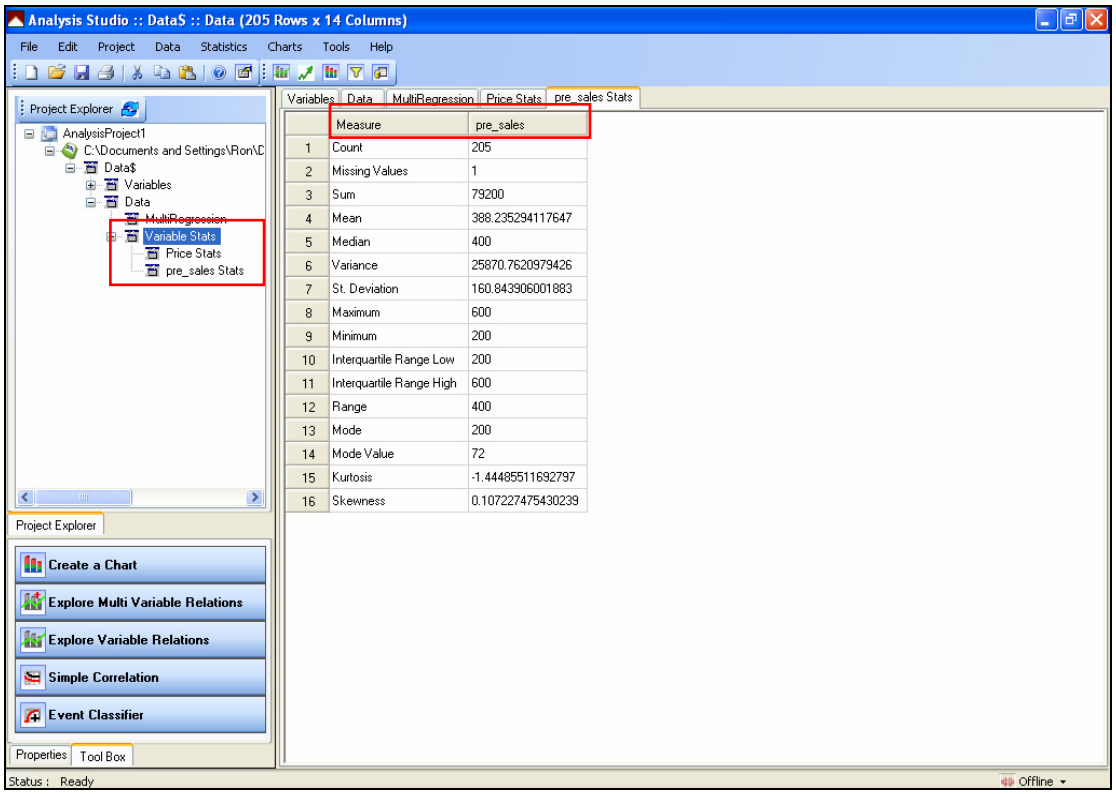
A user can select one or more variables for their parameters. In the above screenshot, the variables "Price" and "pre_sales" are selected. The next screenshot displays the measure values for each selected variable.

Brief Variable Analysis applied to variable results in creation of an entity with the same name accompanied by Stats. This entity appears as a node under the Variable Stats node in the Project Explorer frame. The user can print, close, save or delete each such entity. In addition, each entity appears as a tab in the tab pages frame and when selected its parameters are shown.


Statistical Wizard - Brief Variable Analysis
Click finish to display the results in the output pane.
Add a seperate Grid / Note for each result
Cancel
< Back
Next >
Finish

	Measure	Price	pre_sales
1	Count	205	205
2	Missing Values	1	1
3	Sum	15876	79200
4	Mean	77.8235294117647	388.235294117647
5	Median	79	400
6	Variance	258.707620979427	25870.7620979426
7	St. Deviation	16.0843906001883	160.843906001883
8	Maximum	99	600
9	Minimum	59	200
10	Interquartile Range Low	59	200
11	Interquartile Range High	99	600
12	Range	40	400
13	Mode	59	200
14	Mode Value	72	72
15	Kurtosis	-1.44485511692798	-1.44485511692797

Brief Variable Analysis Tabs and Nodes



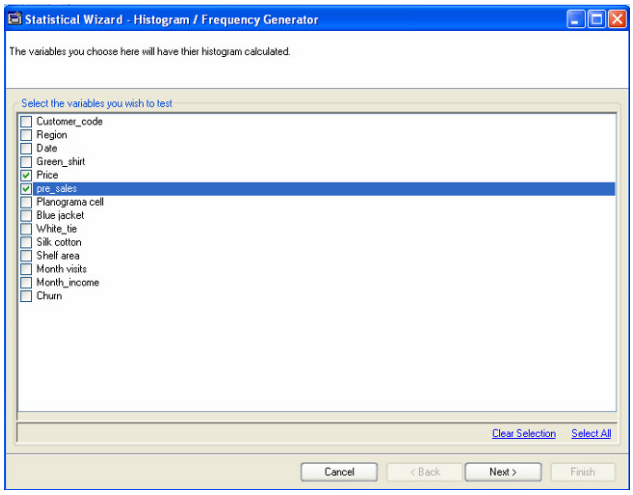
	Measure	pre_sales
1	Count	205
2	Missing Values	1
3	Sum	79200
4	Mean	388.235294117647
5	Median	400
6	Variance	25870.7620979426
7	St. Deviation	160.843906001883
8	Maximum	600
9	Minimum	200
10	Interquartile Range Low	200
11	Interquartile Range High	600
12	Range	400
13	Mode	200
14	Mode Value	72
15	Kurtosis	-1.44485511692797
16	Skewness	0.107227475430239

4.1.5.2 Frequency / Histogram

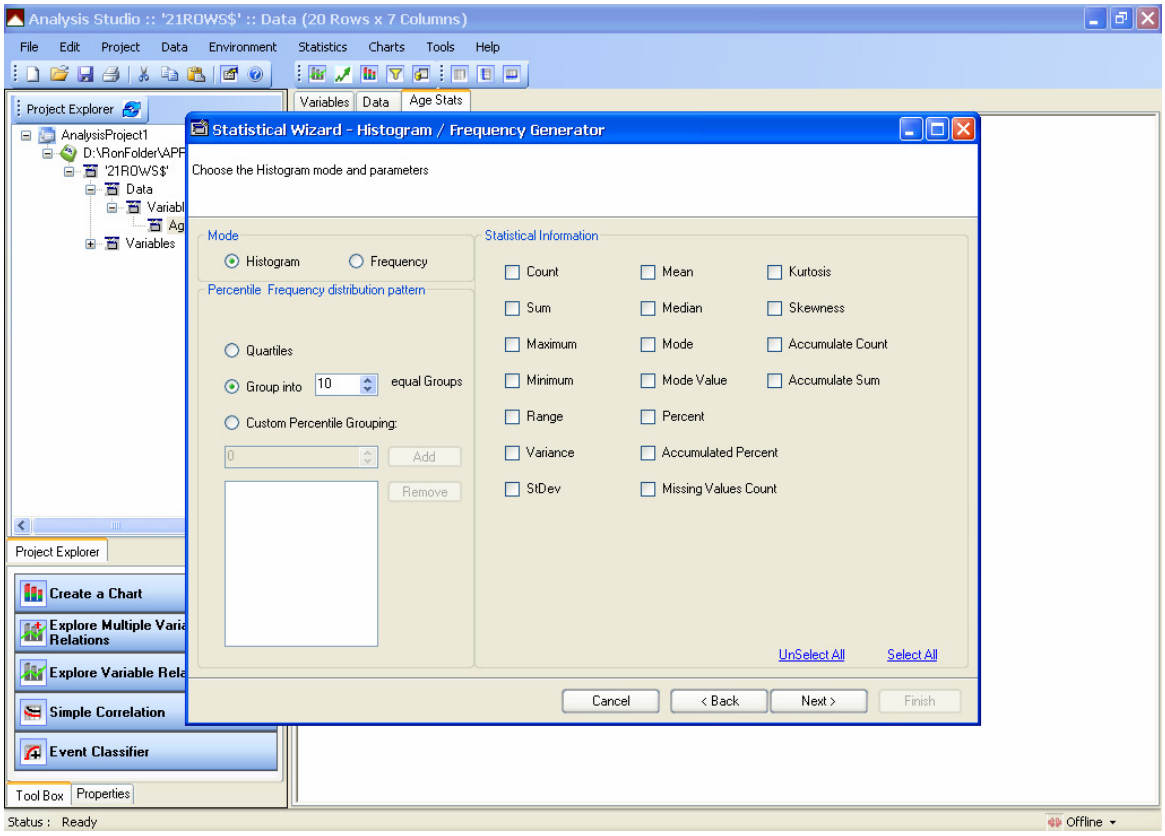
The Frequency / histogram procedure produces frequency and histogram information for obtaining a clear insight into the values of a desired variable. This information is produced according to the selected Histogram mode and parameters.

Frequency / Histogram Wizard Step 1: Selecting Variables

Select “Histogram” mode and parameters for one or more variables. This selection will be applied to all selected variables. It is demonstrated below.

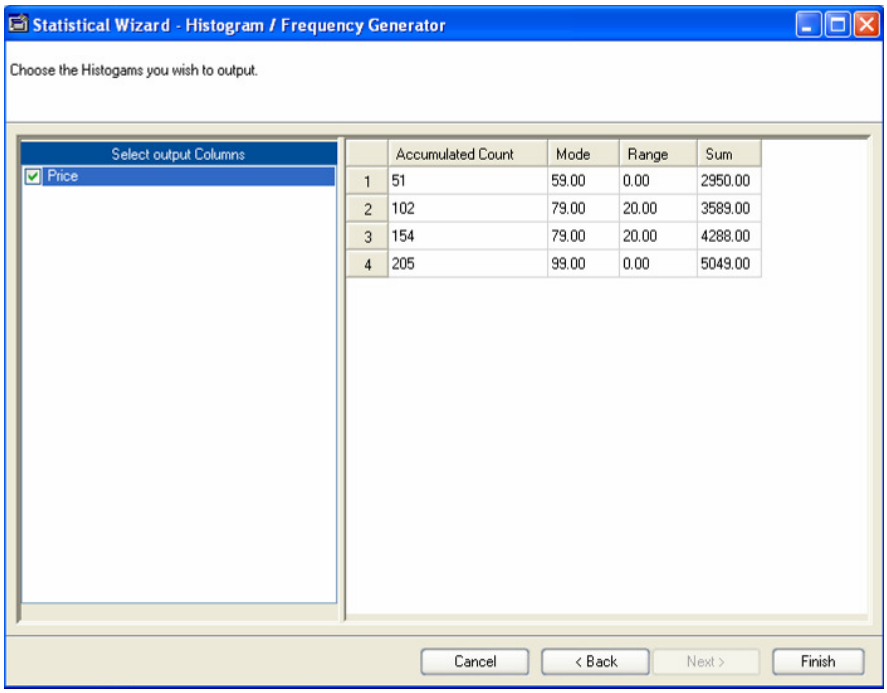


Frequency / Histogram Wizard Step 2: Selecting Parameters



In the above screenshot, the user has selected 4 parameters out of the possible 18 parameters.

Frequency / Histogram Wizard Step 3: Presenting Results



For more than one variable, a user can navigate between variables by clicking on a variable name on the left side of the Wizard window. When a user decides to save the variable information, the checkbox left to its name should be selected.

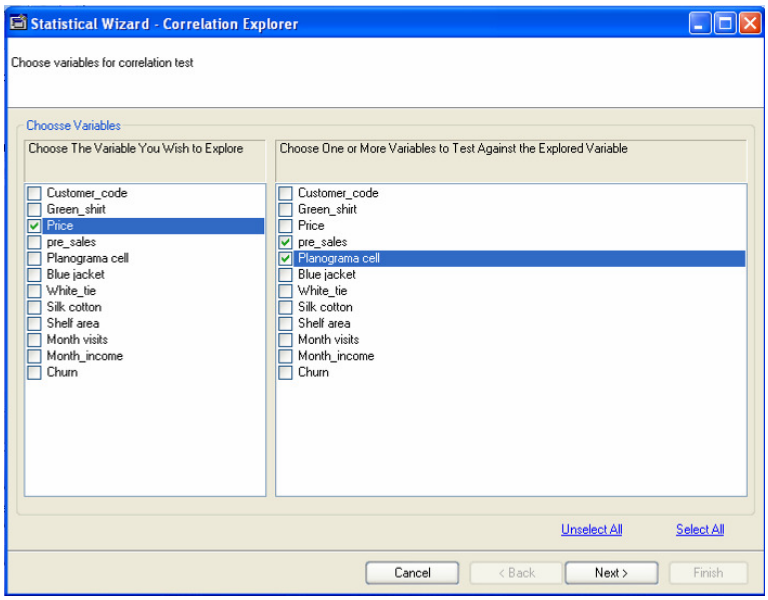
Frequency / histogram applied to a variable results in creation of an entity that appears as a node under the Histograms node in the Project Explorer frame. It has the name of Histogram with a variable name in brackets. The user can print, close, save or delete each such entity. Also, each such entity appears as a tab in the tab pages frame and when selected, it shows its information.

4.1.5.3 Data Correlation

The Data Correlation Wizard computes the Pearson's correlation test that measures the linear strength between two variables. A correlation result is between "-1" to "+1" where "-1" represents a negative perfect linear relation, "0" represents no correlation at all and "+1" represents a positive perfect linear relation. A user should be aware that the two variables have to be normally distributed.

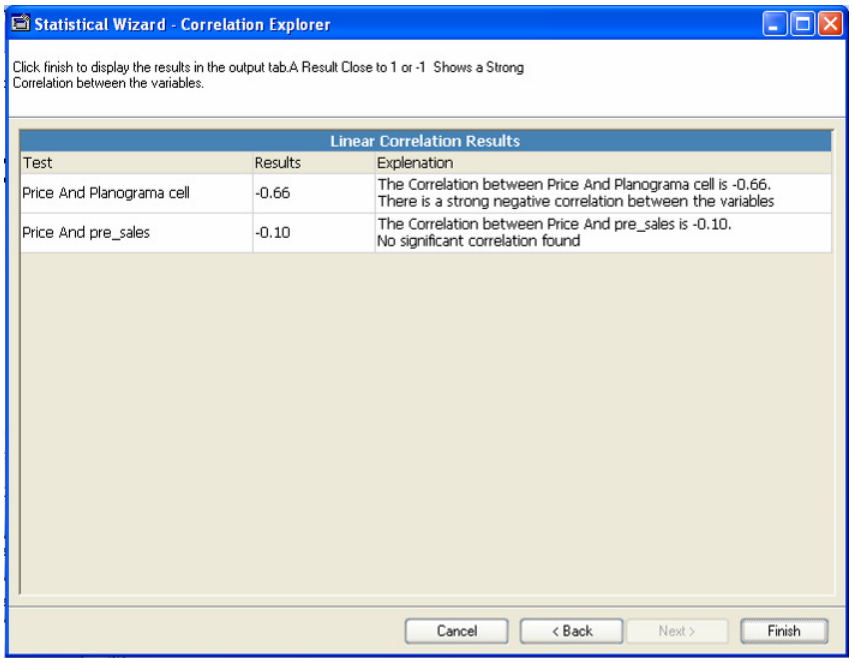
For example: A correlation result of "- 0.9" between "Price" and "Sales level" for product "A" indicates that as "Price" goes up the "Sales level" goes down. A correlation result of "- 0.3" between "Price" and "Sales level" for product "B" indicates that as "Price" goes up the "Sales level" goes down but product "B" has a weaker relation between "Price" and "Sales level" than product "A".

Data Correlation Wizard Step 1: Selecting the Variables



In the above screenshot, two variables ("pre_sales" and "Planograma cell") were selected to be tested for correlation with one variable ("Price").

Data Correlation Wizard Step 2: Presenting Results


Statistical Wizard - Correlation Explorer
Click finish to display the results in the output tab.A Result Close to 1 or -1 Shows a Strong Correlation between the variables.
Linear Correlation Results
Test
Results
Explenation
Price And Planograma cell
-0.66
The Correlation between Price And Planograma cell is -0.66. There is a strong negative correlation between the variables
Price And pre_sales
-0.10
The Correlation between Price And pre_sales is -0.10. No significant correlation found
Cancel
< Back
Next >
Finish

4.1.5.4 The results of the Data Correlation test are presented as an entity on a single tab and as a node on the Project Explorer frame. The user can print, close, save or delete each of the Data Correlation procedure results separately.

4.1.5.5 Auto Data Correlation

The Auto Data Correlation Wizard computes the Pearson's correlation test, which measures the correlation strength between two variables of all or parts of current data set variables. As in the Data Correlation test, a correlation result is between "-1" to "+1" where "-1" represents a negative perfect linear relation, "0" represents no correlation at all and "+1" represents a positive perfect linear relation.

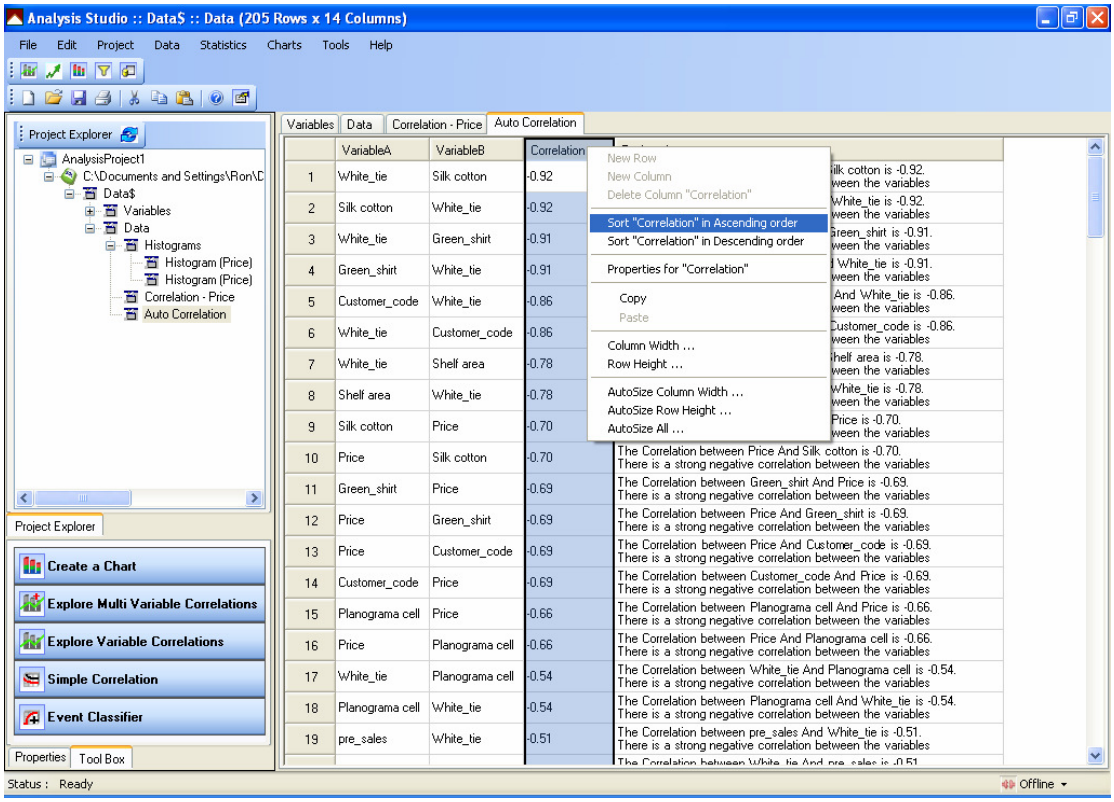
Auto Correlation Test Wizard Step 1: Selecting a Variables

In the above screenshot, all of the current data set variables were selected in order to test the correlation between each other.

Auto Correlation Test Wizard Step 2: Presenting Results

In the above screenshot, the results of all of the correlation between current data set variables are presented as well as their short interpretation. By clicking the “Finish” button the results of the Auto Data Correlation test turn into an entity that is presented as a single tab and a node under the Data node on the Project Explorer frame. A user can print, close, save or delete each such entity separately. The default Auto Correlation name can also be change into a specific one.

4.1.5.6 Data Correlation and Auto Correlation Test Results Display


Analysis Studio :: Data$ :: Data (205 Rows x 14 Columns)
File Edit Project Data Statistics Charts Tools Help
Project Explorer
AnalysisProject1
Data$
Variables
Data
Histograms
Histogram (Price)
Histogram (Price)
Correlation - Price
Auto Correlation
Variables Data Correlation - Price Auto Correlation
VariableA VariableB Correlation
1 White_tie Silk cotton -0.92
2 Silk cotton White_tie -0.92
3 White_tie Green_shirt -0.91
4 Green_shirt White_tie -0.91
5 Customer_code White_tie -0.86
6 White_tie Customer_code -0.86
7 White_tie Shelf area -0.78
8 Shelf area White_tie -0.78
9 Silk cotton Price -0.70
10 Price Silk cotton -0.70
11 Green_shirt Price -0.69
12 Price Green_shirt -0.69
13 Price Customer_code -0.69
14 Customer_code Price -0.69
15 Planograma cell Price -0.66
16 Price Planograma cell -0.66
17 White_tie Planograma cell -0.54
18 Planograma cell White_tie -0.54
19 pre_sales White_tie -0.51
New Row
New Column
Delete Column "Correlation"
Sort "Correlation" in Ascending order
Sort "Correlation" in Descending order
Properties for "Correlation"
Copy
Paste
Column Width ...
Row Height ...
AutoSize Column Width ...
AutoSize Row Height ...
AutoSize All ...
The Correlation between Price And Silk cotton is -0.70.
There is a strong negative correlation between the variables
The Correlation between Green_shirt And Price is -0.69.
There is a strong negative correlation between the variables
The Correlation between Price And Green_shirt is -0.69.
There is a strong negative correlation between the variables
The Correlation between Price And Customer_code is -0.69.
There is a strong negative correlation between the variables
The Correlation between Customer_code And Price is -0.69.
There is a strong negative correlation between the variables
The Correlation between Planograma cell And Price is -0.66.
There is a strong negative correlation between the variables
The Correlation between Price And Planograma cell is -0.66.
There is a strong negative correlation between the variables
The Correlation between White_tie And Planograma cell is -0.54.
There is a strong negative correlation between the variables
The Correlation between Planograma cell And White_tie is -0.54.
There is a strong negative correlation between the variables
The Correlation between pre_sales And White_tie is -0.51.
There is a strong negative correlation between the variables
Create a Chart
Explore Multi Variable Correlations
Explore Variable Correlations
Simple Correlation
Event Classifier
Properties Tool Box
Status : Ready
Offline

4.1.5.7 Simple Regression (Explore Variable Relations)

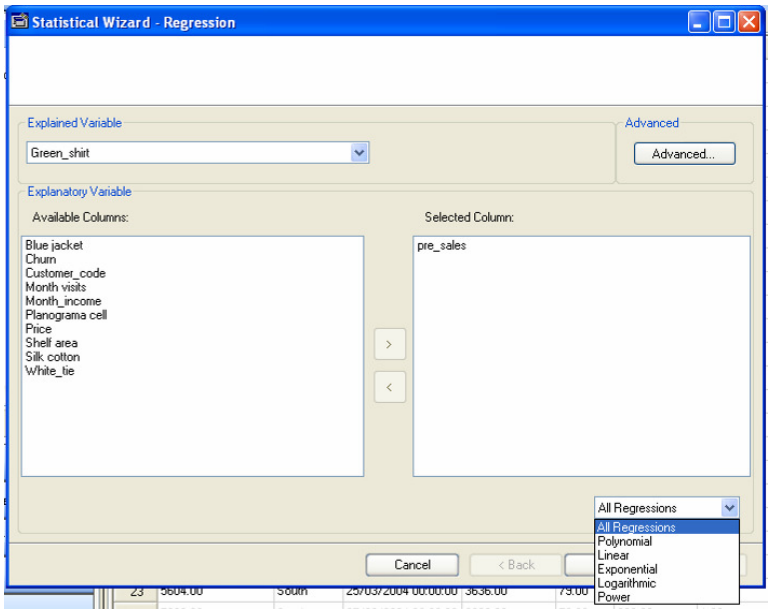
Analysis Studio has a sophisticated and powerful regression analysis procedure. The software supports five regression types, an unlimited number of variables and cases for analysis (limited by computer power only), a unique Best-of-Fit automatic engine, What-If and Sensitivity tables as well as a large number of parameters and charts to assist gaining a good model in a short time.

4.1.5.8 Simple Regression Analysis

Regression is a statistical method used to describe the relationship between two or more variables. In Analysis Studio the simple regression procedure is used to describe the relationship between two variables. Using the regression analysis the user can explore the influence of the variables on each other and predict what will be the variable value by having the other variable value. Unlike correlation tests that require both variables to be normally distributed, in regression analysis only the depended variable has to be normally distributed. The independent variable (x) can have a normal distribution or not.
There are some basic assumptions the user should take into account while performing a simple or multiple-regression analysis. Those assumptions should be considered as assumptions and not as facts since real world simple or multiple regression models rarely contain them all. If the user finds that there is a gross violation of the listed assumptions further steps should be considered.

- The relationships between the explained variable (Y) and explanatory variables (Xi – Xn) are linear.
- For any values of expletory variables (Xi – Xn) the standard deviation of the explained variable (Y) is constant (the same) for all expletory variables (Xi – Xn).
- The explained variable (Y) is normally distributed.
- The errors (Residuals) are independent of probability.

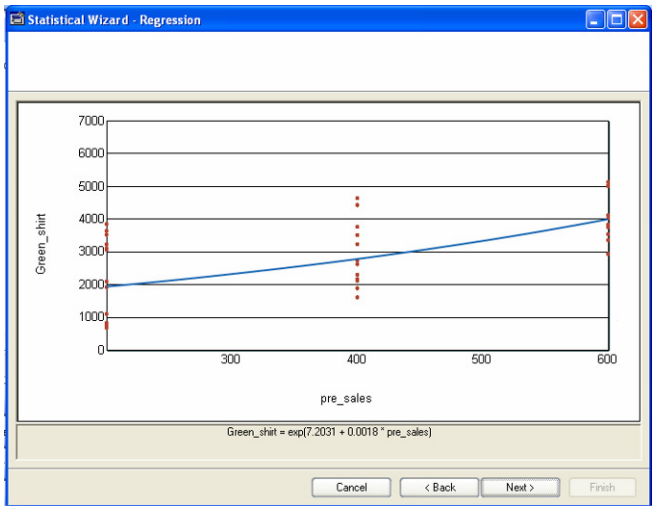
Simple Regression Analysis Wizard Step 1: Selecting the Variables



The first regression analysis screen is used for selecting the desired variables. The user should first select the *Explained Variable* and a single *Explanatory Variable*. By clicking the arrow button the selected explanatory variable will be moved to the S*elected* column.

After selecting the explanatory and the explained variables the user can select one of five regression types or use the default value "All Regressions". If the user selects a specific regression type, the next screen will contain the regression line chart and the regression equation.

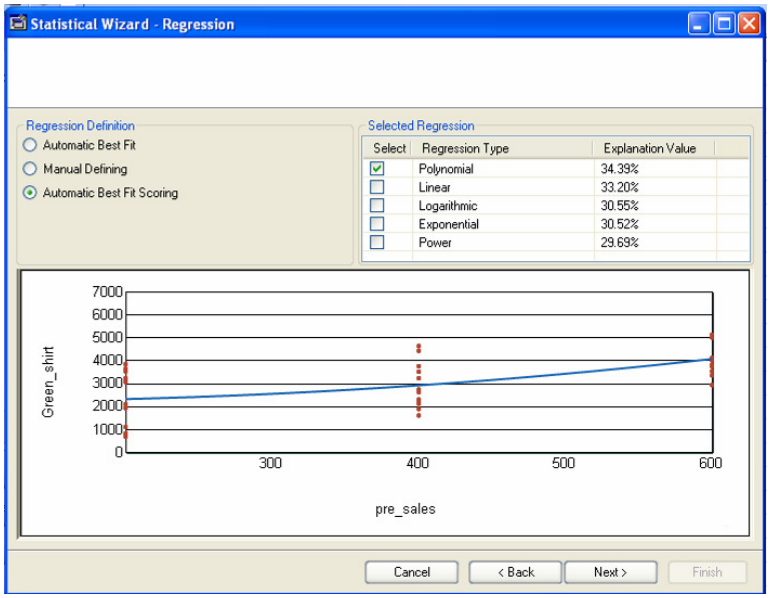
Simple Regression Analysis Wizard Step 2: Displaying a Specific Regression Type Results



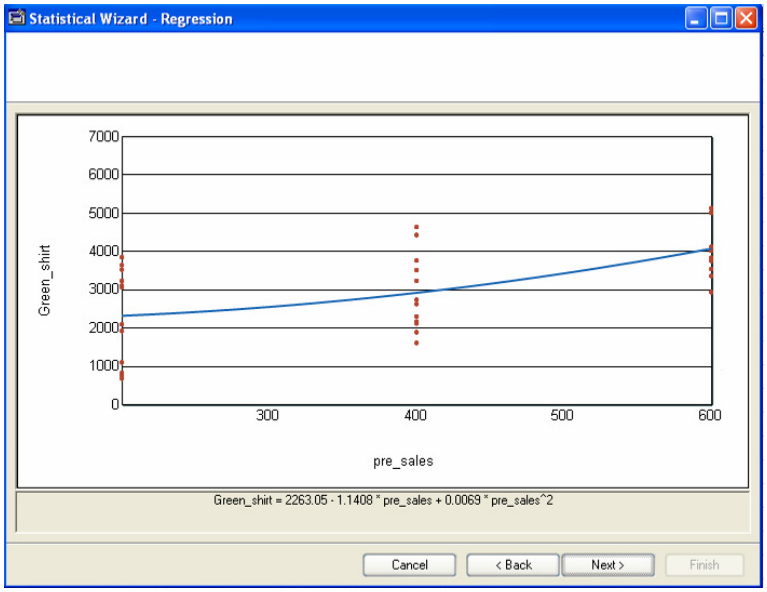
It is advised to make use of the "All Regressions" option because it computes all regression types and allows the user to select the most appropriate and precise type of a given problem. If the user selects the All Regression option, the next screen will contain three display options:

1. **Automatic Best Fit**: presents the regression type that has the best R^2 value.
2. **Manual Defining**: allows the user to determine his preferred regression type.
3. **Automatic Best Fit Scoring**: presents R^2 values from all regressions.

Simple Regression Wizard Step 3: All Regressions Option Results - Automatic Best Fit

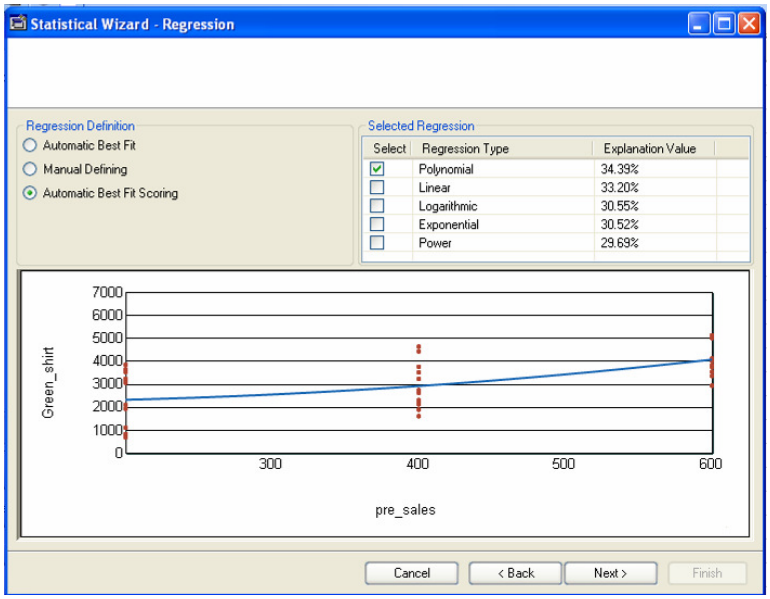


Simple Regression analysis Wizard Step 4: All Regressions option results - Manual Defining

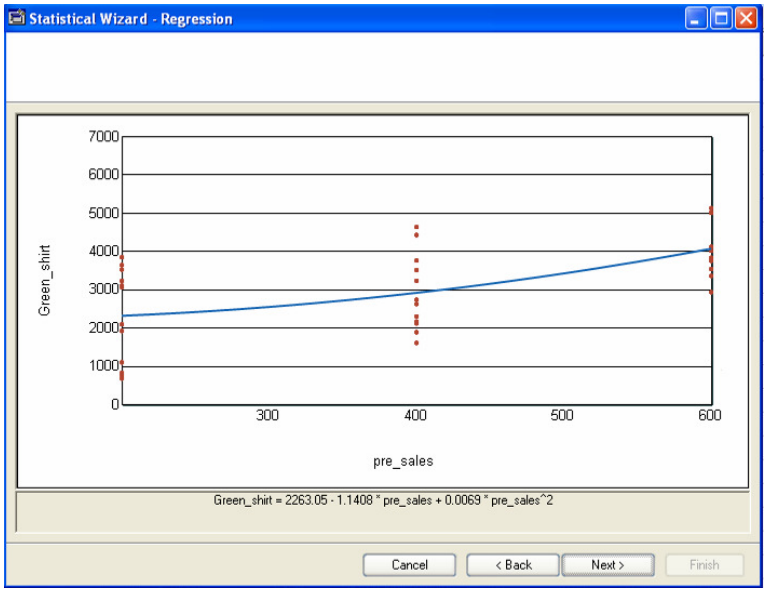


Simple Regression analysis Wizard Step 5: All Regressions Option Results - Automatic Best Fit Scoring

After selecting the desired regression type and clicking the "Next" button a third Wizard screen appears showing the regression chart and equation.

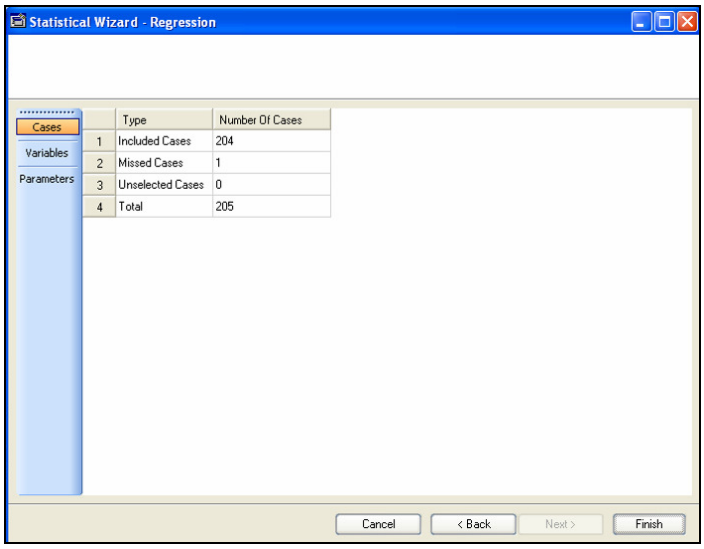


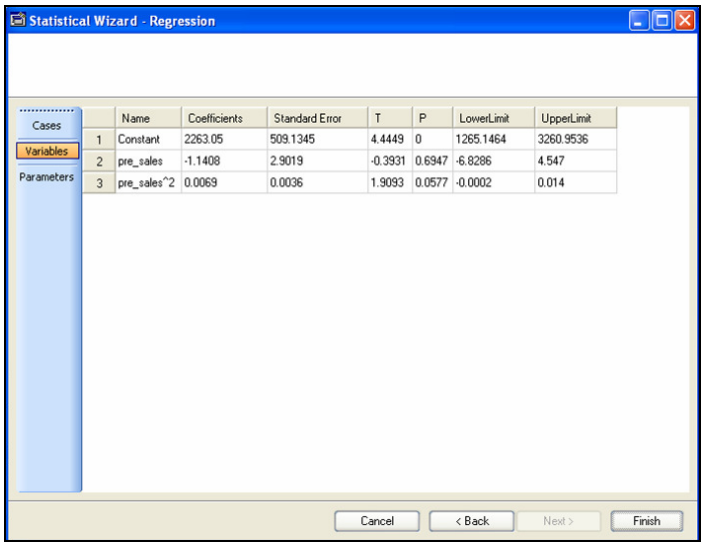
Simple Regression Analysis Wizard Step 6: Displaying Regression Chart and Equation

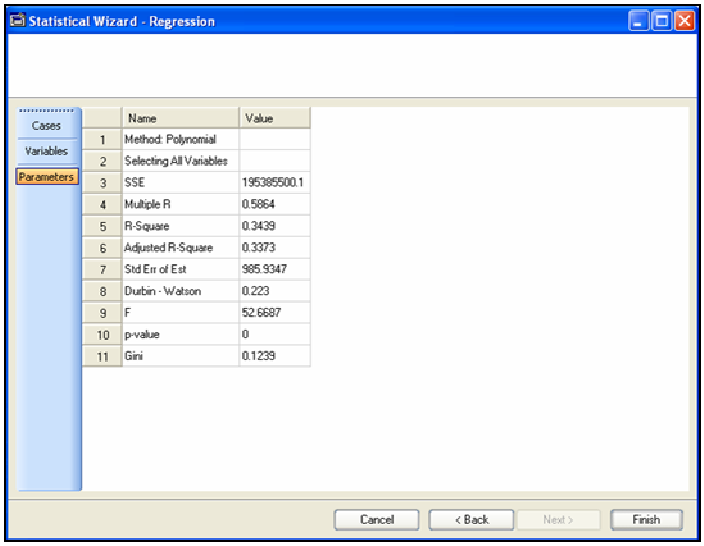


The last Wizard screen is the results screen of the regression computation. The results are presented in the Wizard window to provide the user the opportunity to change the regression type or the data set if the results indicate a problem.

Simple Regression Analysis Wizard Step 7: Displaying Regression Quick Results





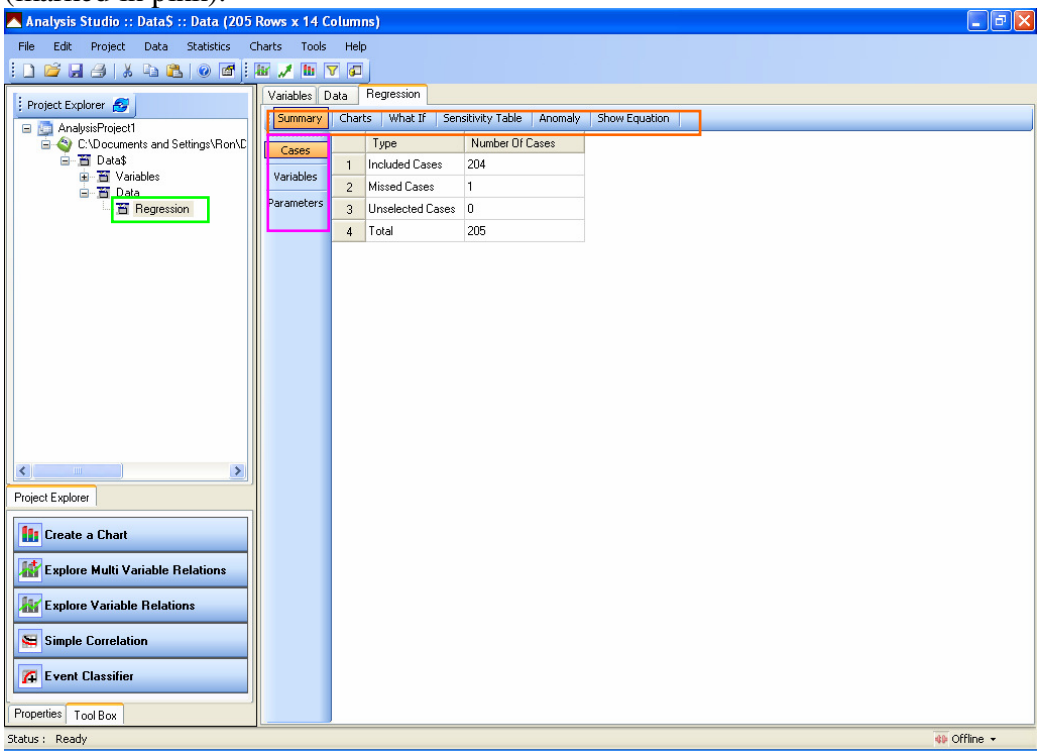


Working with Simple Regression

After the Simple Regression Wizard is complete, the software framework is ready to allow the user to inquire and calculate further, based on the Wizard results.

Simple Regression Framework View

The “Simple Regression” statistical procedure has its own node name and tab window (marked in green for the purpose of this explanation). The Simple Regression framework contains six main frames (marked in orange). Each of them has its own sub-options (marked in pink).



Simple Regression Framework Options:

1. Summary
    a. Cases
    b. Variables
    c. Parameters
2. Charts
    a. Regression
    b. Error
    c. Gain
    d. Lift
3. What-if
    a. What-if Calculator
4. Sensitivity
    a. Sensitivity Calculator
5. Anomaly
    a. Anomaly Calculator
6. Equation

4.1.5.9 Multiple Regression (Exploring Multiple Variable Relations)

In Analysis Studio the multiple regression procedure is used to describe the relationship between two or more variables. There are no limits for the explanatory variables number.

Using multiple regression analysis the user can explore the influence of the explanatory variables on the explained variable and predict what the explained variable value will be by having the other variables value. Unlike a Correlation test that requires that both variables be normally distributed, in regression analysis only the depended variable has to be normally distributed. The independent variables (xn) can have a normal distribution or not. There are some basic assumptions the user should take into account while performing a multiple regression analysis. Those assumptions should be considered as assumptions and not as facts as the real world simple or multiple regression models are rarely containing them all. If the user finds that there is a gross violation of the listed assumptions further steps should be considered.

The relationships between the explained variable (Y) and explanatory variables (Xi – Xn) are linear.

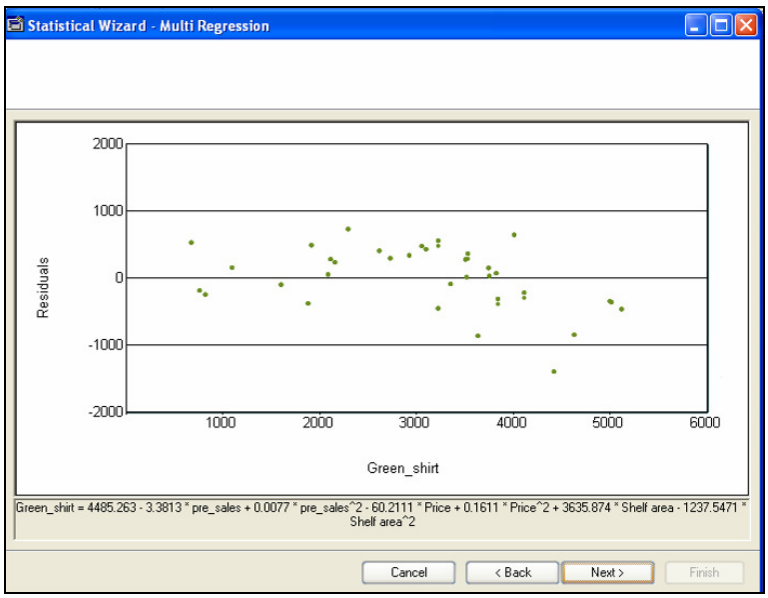
- For any values of expletory variables (Xi – Xn) the standard deviation of the explained variable (Y) is constant (the same) for all expletory variables (Xi – Xn).
- The explained variable (Y) is normally distributed.
- The errors (Residuals) are independent of probability.

Multiple Regression Analysis Wizard Step 1: Selecting the Variables

- The first regression analysis screen is used for selecting the desired variables.
- The user should first select the Explained Variable: to select one or more Explanatory Variables click the arrow button and the selected explanatory variables will be moved to the Selected Columns.
- After selecting the explanatory and the explained variables, the user has two options for gaining a good model:
    1. Selecting multiple regression type.
    2. Selecting best subset method.

1. Selecting Multiple Regression Type:
The user can select one of five multiple regression types or use the default value, "All Regressions". If the user selects a specific regression type, the next screen will contain the multiple regression residuals chart and the regression equation.

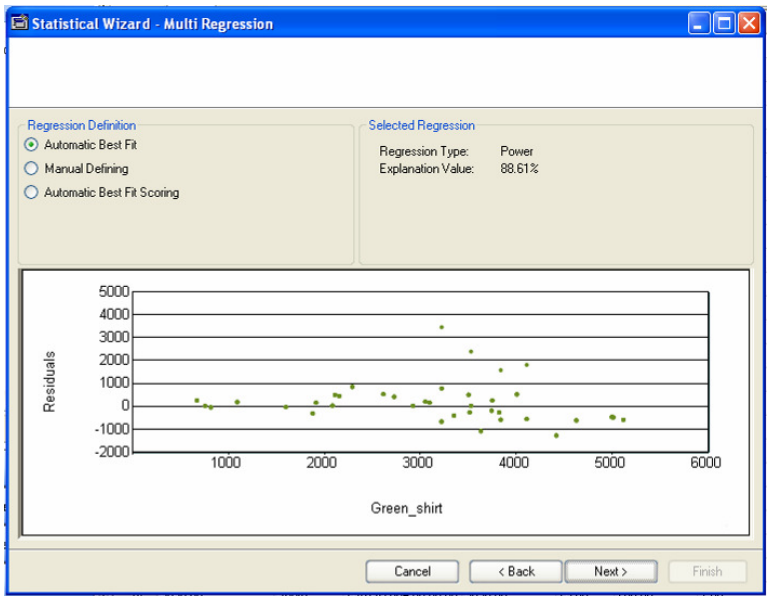
Multiple Regression Analysis Wizard Step 2: Displaying a Specific Regression Type Results (Residuals)



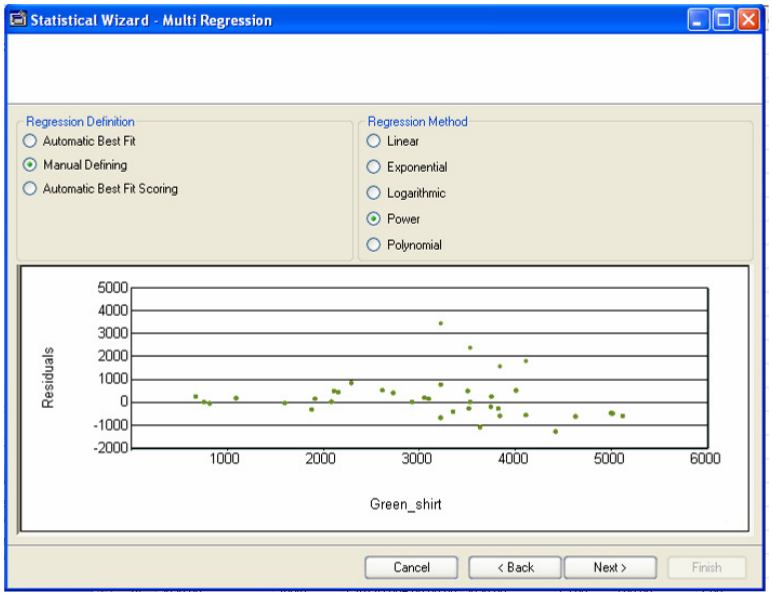
It is advised to make use of the All Regressions option because it computes all regression types and allows the user to select the most appropriate and precise type of a given problem. If the user selects the All Regressions option, the next screen will contain three display options:

1. **Automatic Best Fit**: Presents the regression type that has the best $R^{\wedge 2}$ value.
2. **Manual Defining**: Allows the user to determine his preferred regression type.
3. **Automatic Best Fit Scoring**: Presents $R^{\wedge 2}$ values for all regressions.

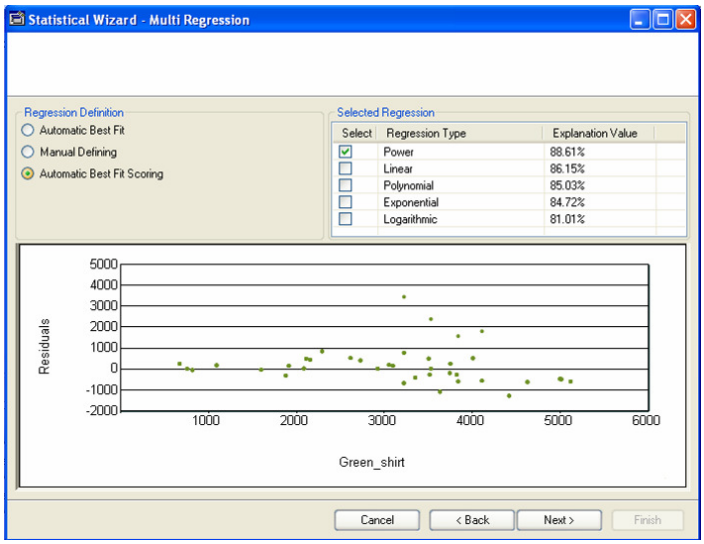
Multiple Regression Analysis Wizard Step 3: All Regressions Option Results, Automatic Best Fit



Multiple Regression Analysis Wizard Step 4: All Regressions Option Results Manual Defining

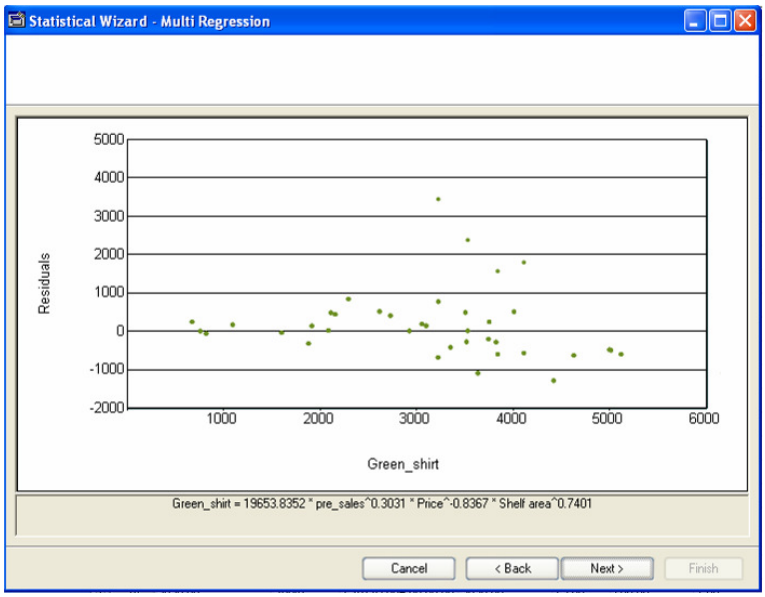


Multiple Regression Analysis Wizard Step 5: All Regressions Option Results, Automatic Best Fit Scoring



After selecting the desired regression type and clicking the "Next" button a third Wizard screen that displays the regression chart and equation appears.

Multiple Regression Analysis Wizard Step 6: Displaying Regression Chart and Equation



2. Selecting Best Subset Method:

For multiple regression analysis that contains a large number of explanatory variables a method for reducing the explanatory variables number is required in order to have a stable model that is easy to interpret.

By using a best subset method, the user can avoid over-fitting an unnecessary complex model.

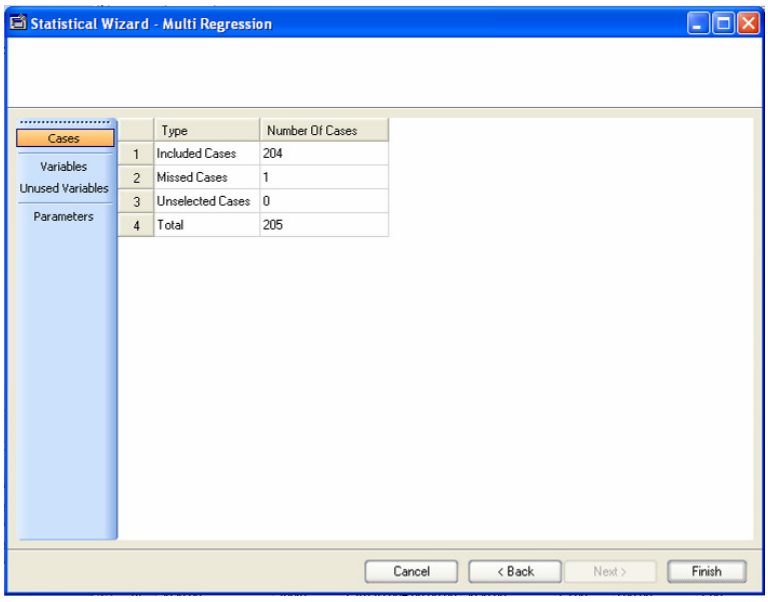
Analysis Studio default best subset method is *Enter All* which means that the software computes all variables excluding only inner-linear correlation variables. The software supports two additional best subset methods:

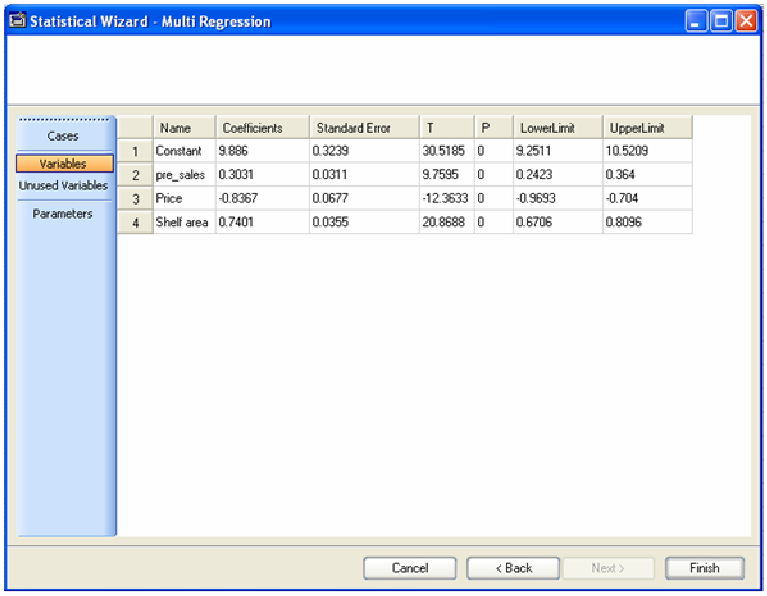
1. **Stepwise by P value**
2. **Stepwise by Adjusted R squared**

The user can select the desired method based on personal preference.

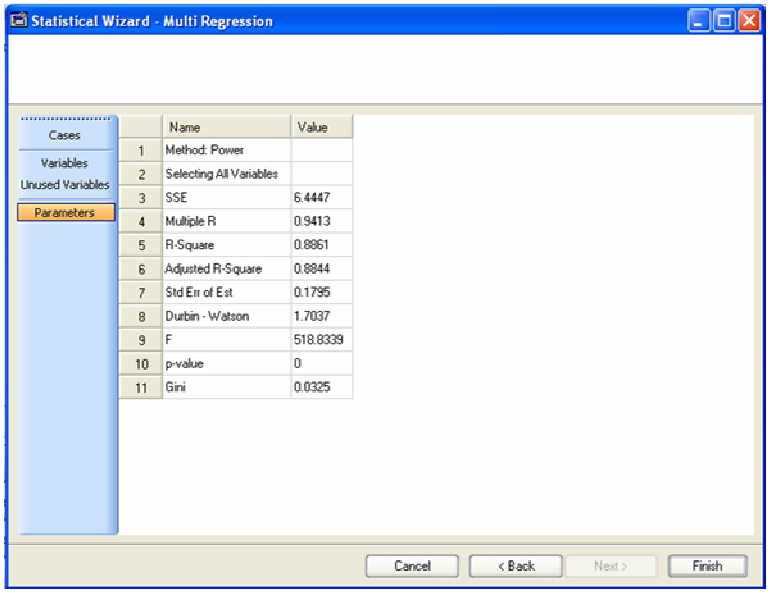
The last Wizard screen is the results screen of the regression computation.
The results are presented in the Wizard to provide the user with the opportunity of changing the regression type or the data set if the results indicate a problem.

Multiple Regression Analysis Wizard Step 7: Displaying Regression Quick Results





	Name	Coefficients	Standard Error	T	P	LowerLimit	UpperLimit
1	Constant	9.886	0.3239	30.5185	0	9.2511	10.5209
2	pre_sales	0.3031	0.0311	9.7595	0	0.2423	0.364
3	Price	-0.8367	0.0677	-12.3633	0	-0.9693	-0.704
4	Shelf area	0.7401	0.0355	20.8688	0	0.6706	0.8096

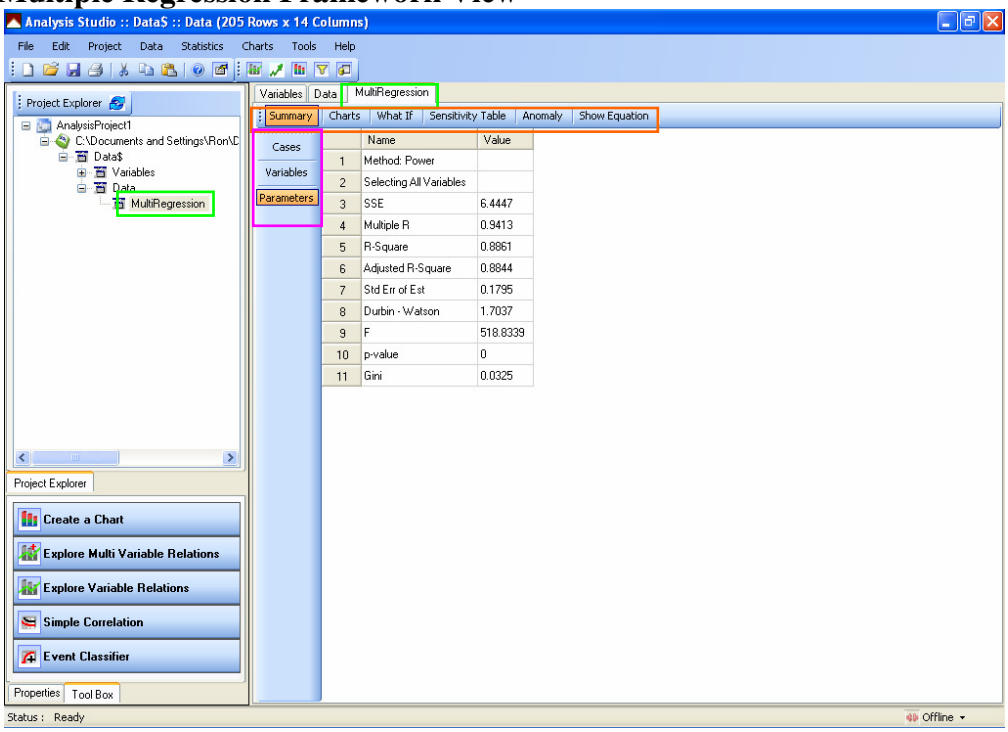


	Name	Value
1	Method: Power	
2	Selecting All Variables	
3	SSE	6.4447
4	Multiple R	0.9413
5	R-Square	0.8861
6	Adjusted R-Square	0.8844
7	Std Err of Est	0.1795
8	Durbin - Watson	1.7037
9	F	518.8339
10	p-value	0
11	Gini	0.0325

Working with Multiple Regression

After finishing the multiple regression Wizard, the software framework is ready to allow the user to inquire and calculate further, based on the Wizard results.

Multiple Regression Framework View



As any other statistical procedure the multiple regression procedure has its own node name and tab window (marked in green for the purpose of this explanation). The multiple regression framework contains six main frames (marked in orange). Each of them has its own sub options (marked in pink).

Multiple Regression Framework Options:

1. Summary
    a. Cases
    b. Variables
    c. Parameters
2. Charts
    a. Residuals
    b. Error
    c. Gain
    d. Lift
3. What-if
    a. Multiple variables What-if calculator
4. Sensitivity
    a. Multiple variables Sensitivity calculator
5. Anomaly
    a. Multiple variables anomaly calculator
6. Equation display

Logistic Regression

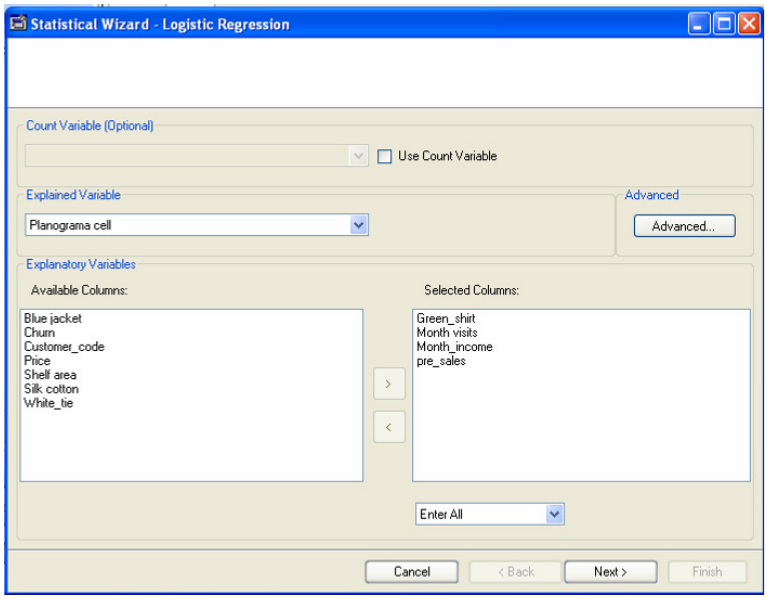
The goal of Logistic Regression is to find the best fitting model that describes the relationship between an explained variable and one or more explanatory variables. The explained variable in logistic regression is binary (i.e. smoker or nonsmoker, churn or non-churn etc.). The program includes cases with "0" for False and "1" for True. Cases with values other then 0 or 1 for the binary that is the explained variable will be excluded from the model.

Analysis Studio treats logistic regression in a holistic point of view meaning that the building of the model and the interpretation of the results are aimed to assist the user in finding the best fitting model and in having the ability to perform interactive simulation based on the model.

Logistic Regression Wizard

Analysis Studio interactive logistic regression Wizard has three screens that support an in-depth classification analysis. The user can change the model settings during the model building.

Logistic Regression Wizard Step 1: Building the Logistic Regression Model



Steps for default logistic regression modeling are as followed:

1. The program automatically detects binary variables and places them at the Explained variable combo box. The user can choose the desired explained binary variable from the combo box. The explanatory variables are displayed in the right list box.
2. The user marks the desired variables and by clicking on the arrow button, the desired variables will be selected for the logistic regression procedure.

The program has four modeling methods for the logistic regression:

1. **Enter all**: all variables that are marked as "selected columns" will participate in building the model unless the program detects that one or more of the variables have inner correlations with the explained variable. Variables that have an inner correlation will be excluded from the model.

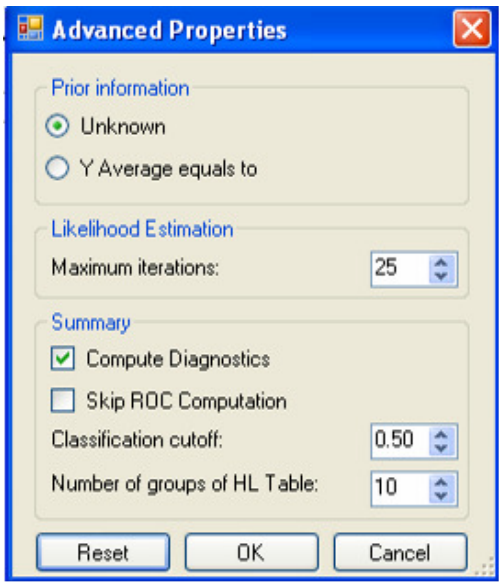
2. **Stepwise Selection by P-Value**: the program computes the significance of each additional variable sequentially; after entering a variable in the model, it checks and removes variables that became insignificant to the model's performance.

3. **Stepwise Selection by AIC**: the software computes the Akaike Information Criterion. It quantifies the relative goodness-of-fit of various previously derived statistical models, given a sample of data. The driving idea behind the AIC is to examine the complexity of the model together with how well it fits with the sample data, and to produce a measure that balances between the two. The selected subset will be the subset that an additional variable cannot improve the logistic regression model. If the model goal is to perform as good as possible prediction, then the AIC can be the right choice. If the user compares two or more prediction models the model with the lowest AIC value is considered to be better.

4. **Stepwise Selection by SIC**: the software computes the Schwarz Information Criterion Information. The SIC is an alternative to the AIC and considered more stable but with lower prediction performances. If the user compares two or more prediction models the model with the lowest SIC value is considered to be better.

Advance Button

The advance button opens a set of advanced properties for logistic regression modeling.



Prior Information

The “Prior Information” sub-frame is for recalculating the explained variable (Y) if there is prior knowledge of the explained variable (Y) rate in the data set.

Likelihood Estimation

The user can determine the number of maximum iterations; as the number of maximum iterations grows so does the accuracy of parameters, but usually in a minor way. The default value is sufficient for gaining good accuracy while increasing the maximum iterations can slow down the model building time.

Summary

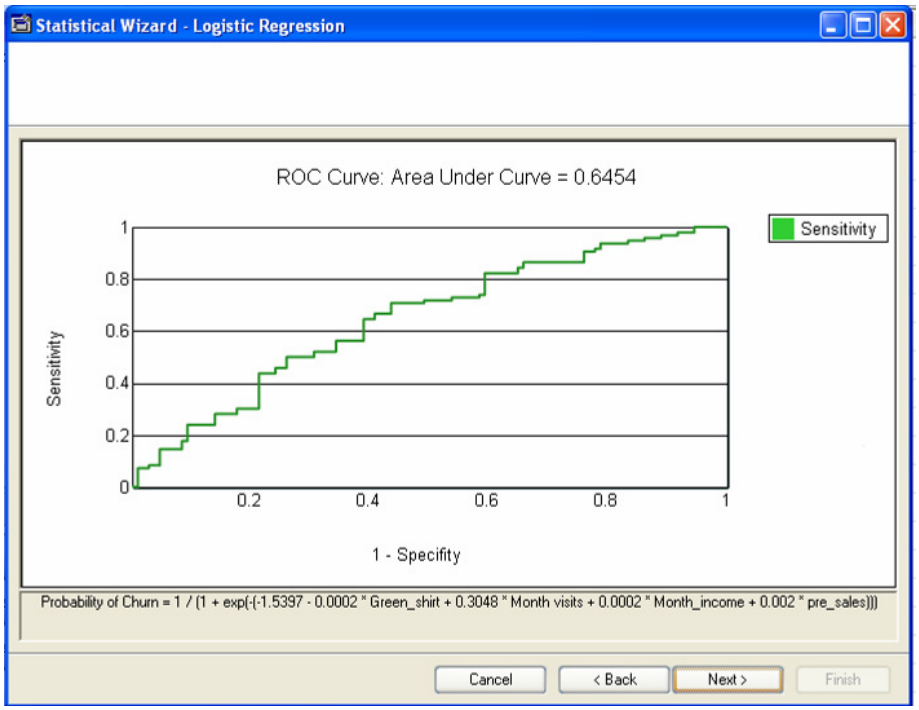
Compute Diagnostics
The logistic regression procedure has diagnostics computations including charts. Such computation can be time consuming, so an option to skip this step is available. The default option is to have the diagnostics computations done and displayed.

Skip ROC Computation
ROC is considered as an additional computation to logistic regression. The software allows the user to determine whether to include the ROC computation as part of the model building.

Classification Cutoff
This option gives the user the ability to change the cutoff value of the predicted explained variable (Y) while building the logistic regression model. The default value of 0.5 is the commonly used default value. It is recommended that only well trained users change these settings.

Number of Group of HL Table
This option allows the user to change the commonly used default value of 10 groups. It is recommended that only well trained users change these settings.

Logistic Regression Wizard Step 2: Viewing the ROC Parameter and Model Equation



In the screenshot above, the ROC computation is presented along with the model equation. Based on these first results the user can decide whether to proceed with the proposed model or compute another one in order to potentially gain better results.

ROC

ROC is used to measure a model's ability to distinguish between two values (i.e. the ability to distinguish between churn customer and non-churn customer). The ROC is the curve itself, while AUC refers to the "Area Under Curve".

- If AUC = 0.5: If the Area Under Curve is 0.5 this means that the model cannot distinguish between the two values more than a random guess.

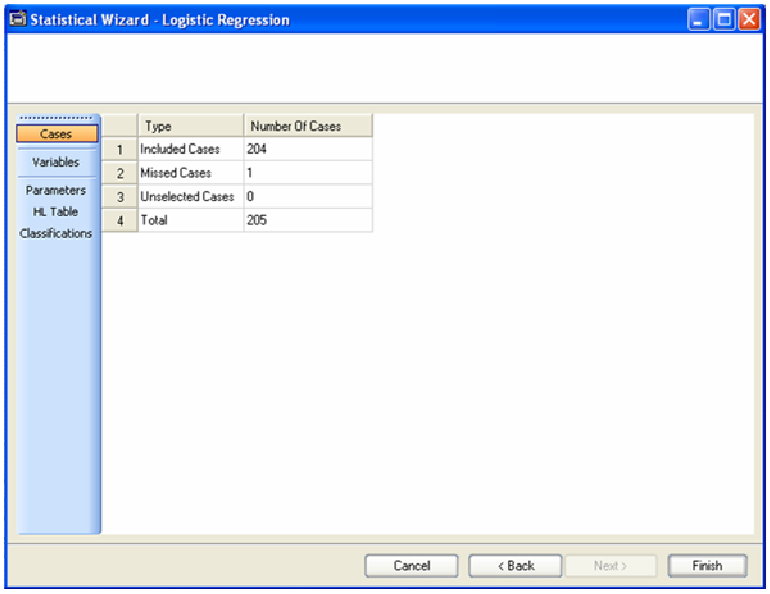
- If AUC < ROC < 0.9: If the Area Under Curve of the model exceeds 0.5 – 0.9 this means that the model can successfully distinguish between the two values.

- If AUC > 0.9: If the Area Under Curve exceeds 0.9 it is advised to check the model variables or to check for over fitting.

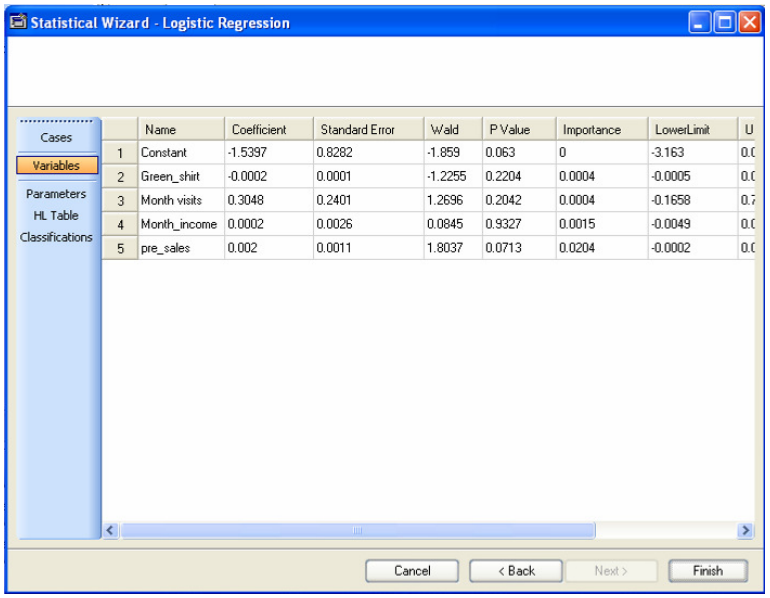
Most business models have AUC of 0.70-0.85.

Logistic Regression Analysis Wizard Step 3: Displaying Regression Quick Results

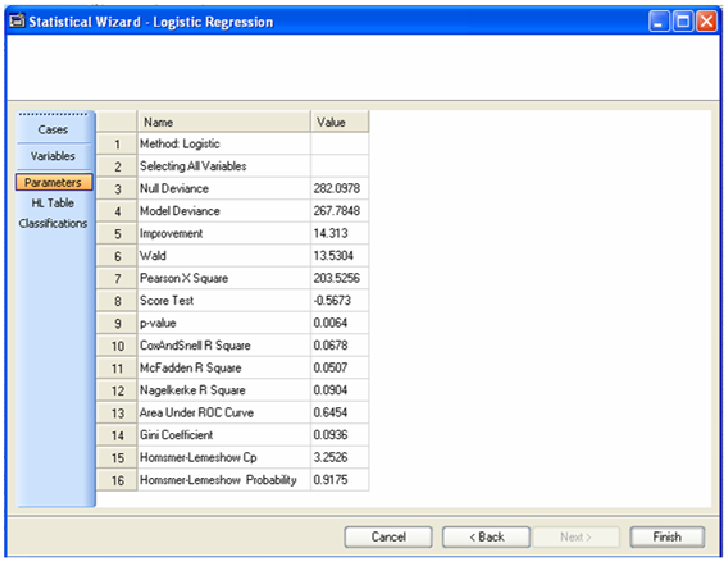
Cases Summary Overview Screen:



Variables Summary Overview Screen:

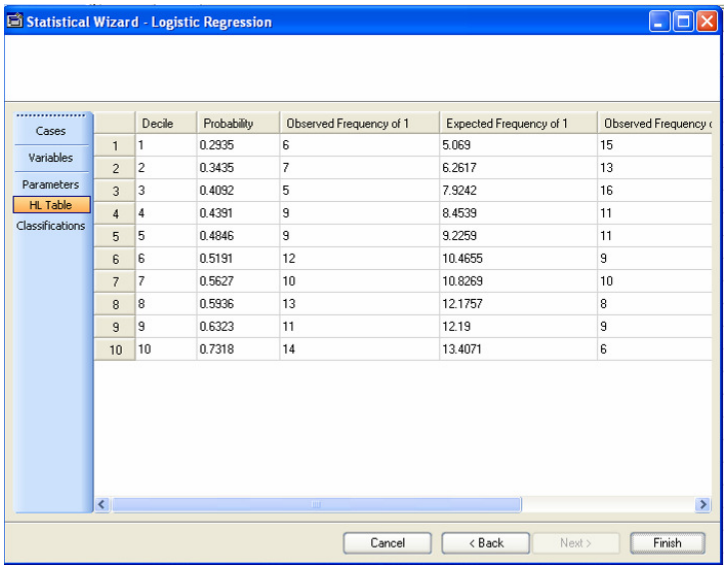


Logistic Regression Output Summary Screen



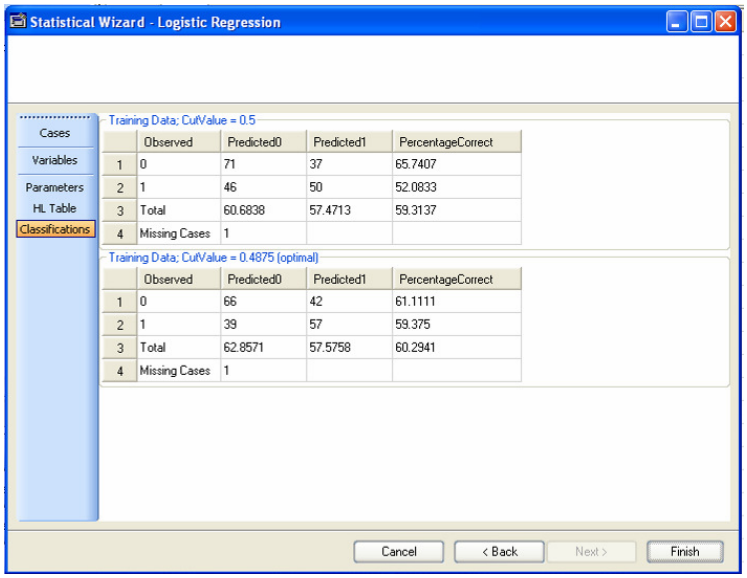
	Name	Value
1	Method: Logistic	
2	Selecting All Variables	
3	Null Deviance	282.0978
4	Model Deviance	267.7848
5	Improvement	14.313
6	Wald	13.5304
7	Pearson X Square	203.5256
8	Score Test	-0.5673
9	p-value	0.0064
10	CoxAndSnell R Square	0.0678
11	McFadden R Square	0.0507
12	Nagelkerke R Square	0.0904
13	Area Under ROC Curve	0.6454
14	Gini Coefficient	0.0936
15	Hornsmer-Lemeshow Cp	3.2526
16	Hornsmer-Lemeshow Probability	0.9175

Logistic Regression Hosmer & Lemeshow Table Screen



	Decile	Probability	Observed Frequency of 1	Expected Frequency of 1	Observed Frequency
1	1	0.2935	6	5.069	15
2	2	0.3435	7	6.2617	13
3	3	0.4092	5	7.9242	16
4	4	0.4391	9	8.4539	11
5	5	0.4846	9	9.2259	11
6	6	0.5191	12	10.4655	9
7	7	0.5627	10	10.8269	10
8	8	0.5936	13	12.1757	8
9	9	0.6323	11	12.19	9
10	10	0.7318	14	13.4071	6

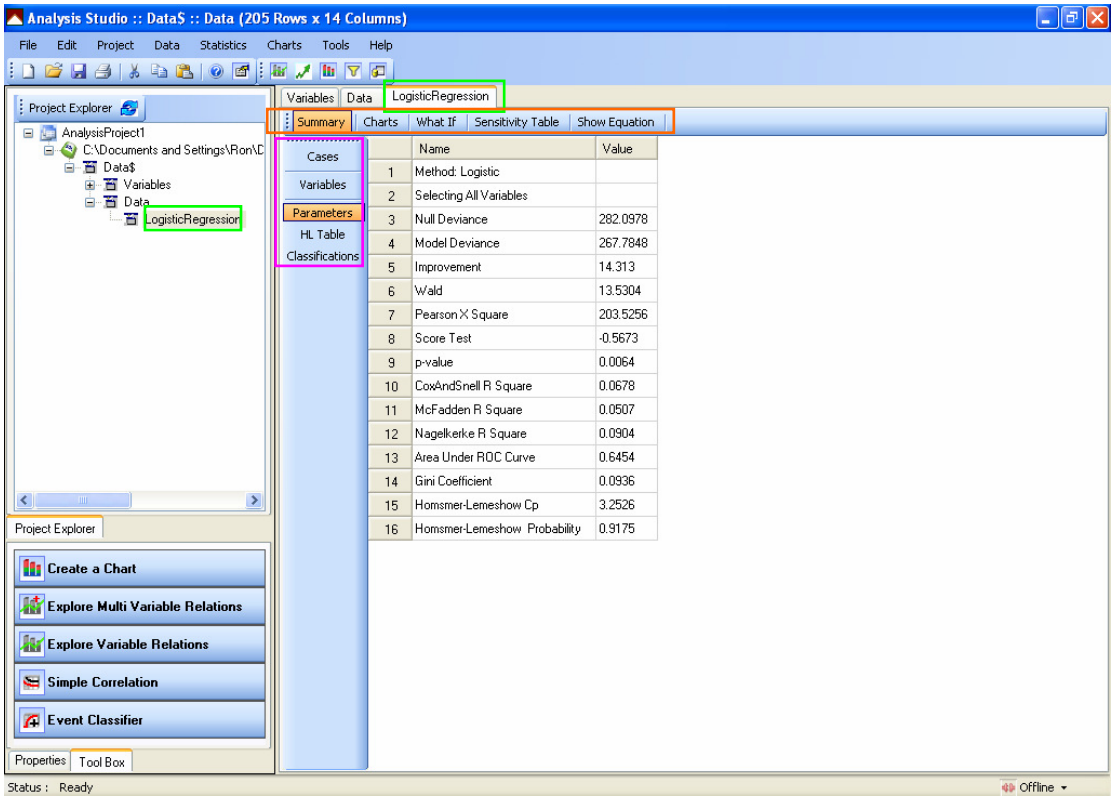
Logistic Regression Classification Optimization Screen



Working with Logistic Regression

After finishing the Logistic Regression Wizard, the software framework is ready to allow the user to inquire and calculate additional regressions based on the Wizard results.

Logistic Regression Framework View



The Logistic Regression as any other statistical procedure has its own node name and tab (marked by a green frame for the purpose of this explanation). The Logistic Regression framework contains five main frames (marked by orange). Each of them has its own sub options (marked by pink).

Logistic Regression Framework Options:

1. Summary
    a. Cases
    b. Variables
    c. Parameters
    d. HL Table
    e. Classification
2. Charts
    a. ROC
    b. Cut points
    c. Gain
    d. Lift
    e. X Diagnostics
    f. Y Diagnostics
    g. Cases
    h. Hits/Misses
    i. Hits Ratio
    j. Misses Ratio
3. What-if
    a. Multiple variables What-if calculator
4. Sensitivity
    b. Multiple variables Sensitivity calculator
5. Equation display

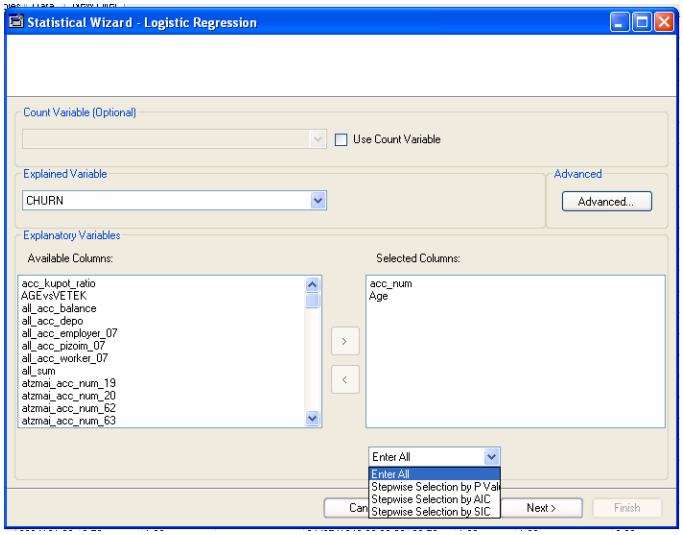
4.15.11 Logistic Regression Fractional Polynomials (F.P.)

Fractional Polynomials is a method that can help create a better logistic regression model by producing 36 different types of transformations for each Continuous variable that is accounted for by the model. By deploying Polynomial transformations the probability that some of the transformations will be more suitable than the original variable for the model increases. Similar to the Logistic regression the program has three modeling methods for the logistic regression that are based on Fractional Polynomials:

Logistic Regression (F.P.) Wizard

Analysis Studio interactive logistic regression Wizard for F.P. calculations has four screens that support an in-depth classification analysis. The user can change the model settings during the model building.

Logistic Regression (F.P.) Wizard Step 1: Building the Logistic Regression (F.P.) Model



Steps for default logistic regression (F.P.) modeling are as follows:

1. The program automatically detects binary variables and places them at the Explained variable combo box. The user can choose the desired explained binary variable from the combo box.

2. The explanatory variables are displayed at the right list box. The user marks the desired variables and by clicking on the arrow button, the desired variables will be selected for the logistic regression (F.P.) procedure.

The program has four modeling methods for the logistic regression:

1. **Enter all**: all variables that are marked as "selected columns" will participate in building the model unless the program detects that one or more of the variables have inner correlations with the explained variable .Variables that have inner correlation will be excluded from the model.

2. **Stepwise selection by p-value**: the program computes the significance of each additional variable sequentially; after entering a variable in the model, it checks and removes variables that become insignificant to the model's performance.

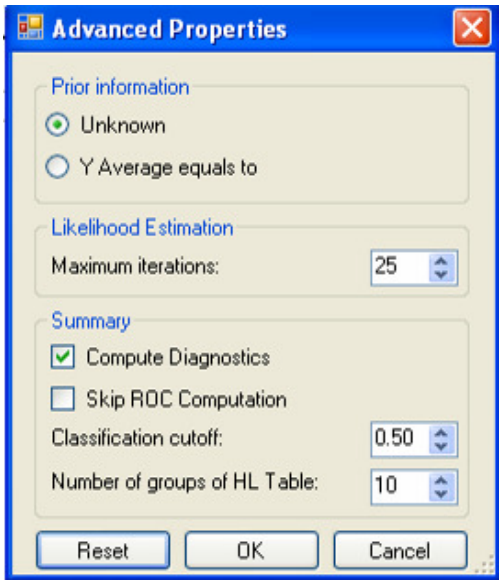
3. **Stepwise selection by AIC**: the software computes the Akaike Information Criterion. It quantifies the relative goodness-of-fit of various previously derived statistical models, given a sample of data. The driving idea behind the AIC is to examine the complexity of the model together with how well it fits with the sample data, and to produce a measure that balances between the two. The selected subset will be the subset that an additional variable cannot improve the logistic regression model. If the goal of the model is to perform an "as good as possible" prediction, then the AIC can be the right choice. If the user compares two or more prediction models the model with the lowest AIC value is considered to be better.

4. **Stepwise Selection by SIC**: the software computes the Schwarz Information Criterion Information. The SIC is an alternative to the AIC and considered more

stable but with lower prediction performances. If the user compares two or more prediction models the model with the lowest SIC value is considered to be better.

Advance Button

The advance button opens a set of advanced properties for logistic regression (F.P.) modeling.



Prior Information

The sub-frame Prior Information is for recalculation of the explained variable (Y) if there is prior knowledge of the explained variable (Y) rate in the data set.

Likelihood Estimation

The user can determine the number of maximum iterations; as the number of maximum iterations grows so does the accuracy of parameters, but usually in a minor way. The default value is sufficient for gaining good accuracy while increasing the maximum iterations can slow down the model building time.

Summary

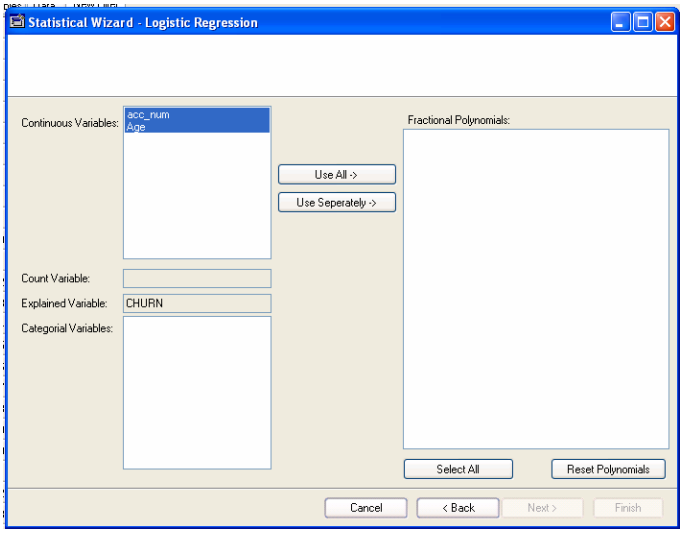
Compute Diagnostics
The logistic regression procedure has diagnostics computations including charts. Such computation can be time consuming, so an option to skip this step is available. The default option is to have the diagnostics computations done and displayed.

Skip ROC Computation
ROC is considered as an additional computation to logistic regression. The software allows the user to determine whether to include the ROC computation as part of the model building.

Classification Cutoff
This option gives the user the ability to change the cutoff value of the predicted explained variable (Y) while building the logistic regression model. The default value of 0.5 is the commonly used default value. It is recommended that only well trained users change these settings.

Number of Group of HL Table
This option allows the user to change the commonly used default value of 10 groups.
It is recommended that only well trained users change these settings.

Logistic Regression (F.P.) Wizard Step 2: Selecting the Desired Variables for the Model

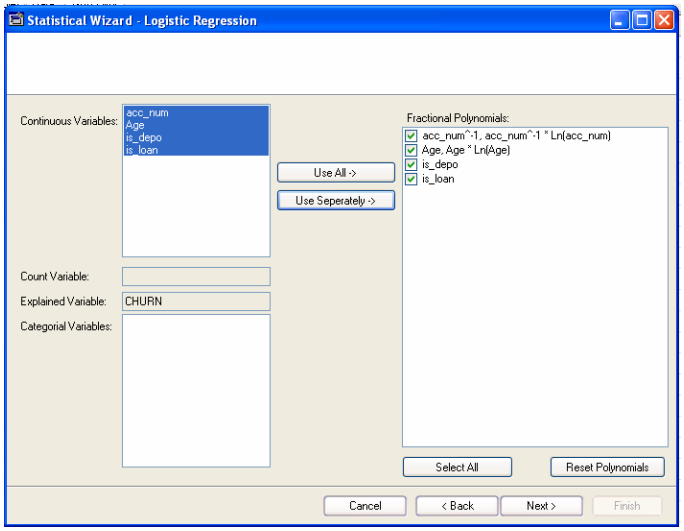


Selecting the desired variables for the F.P. calculations can be done by selecting them on the top left frame. In the above example two continuous variables are selected. If the variables set contains categorical variables they will appear in the left frame but the F.P. calculation will not include them until the final model building.

Use all option
By clicking the *Use All* option the F.P. model will use all variables in the F.P. calculation. All continuous variables will have fractional polynomials calculations upon them.

Use Separately
By clicking the *Use Separately* option the F.P. model will use the selected variables only in the F.P. calculation. All continuous variables will have fractional polynomials calculations upon them.

Both options are the trigger for deploying the F.P. calculation. As the F.P. calculations are done the variable transformations are presented on the right frame for user inspection.
The user can control which variables will be selected to the final logistic regression model by clicking on the *check boxes* of the variables. The option to reset all Polynomials is also available.



In the screenshot above the model contains polynomials transformations for two continuous variables and two categorical variables. All four variables will enter the logistic regression model according to the four modeling methods of the logistic regression.

ROC

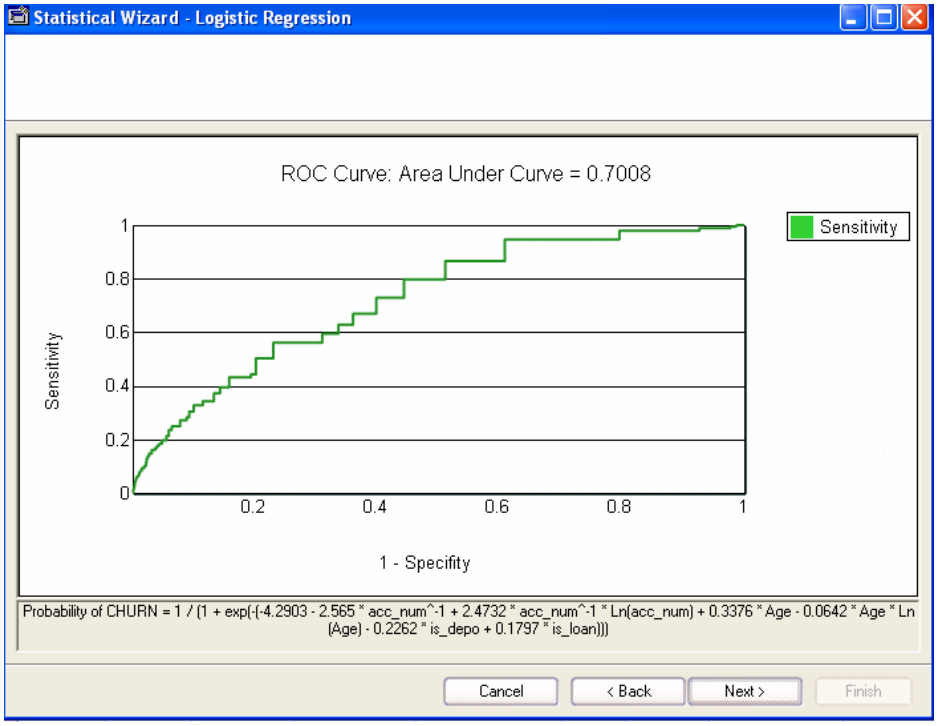
ROC is used to measure a model's ability to distinguish between two values (i.e. the ability to distinguish between churn customer and non-churn customer). The ROC is the curve itself, while AUC refers to the "Area Under Curve".

- If AUC = 0.5: If the Area Under Curve is 0.5 this means that the model cannot distinguish between the two values more than a random guess.

- If 0.5 < AUC < 0.9: If the Area Under Curve of the model exceeds 0.5 – 0.9 this means that the model can successfully distinguish between the two values.

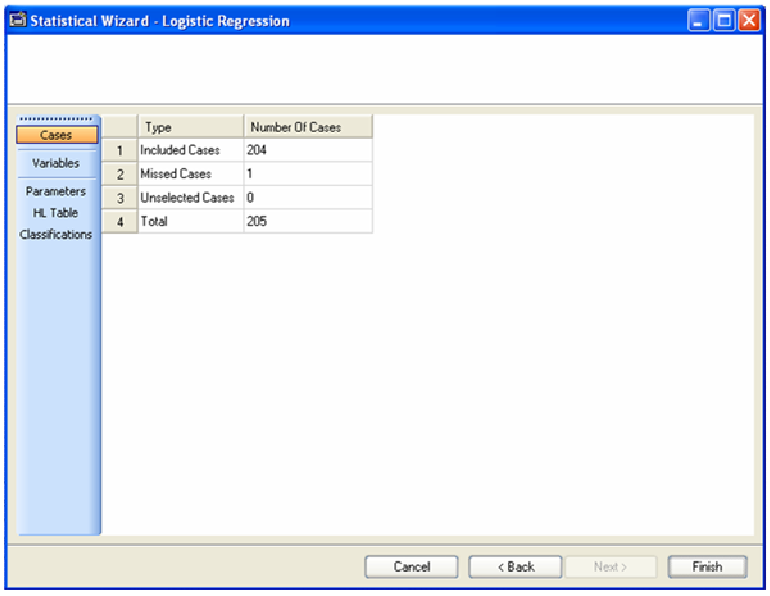
- If AUC > 0.9: If the Area Under Curve exceeds 0.9 it is advised to check the model variables or to check for over fitting.

The ROC has the same interpretation when the F.P. method is performed as in the Logistic regression model. In most cases the ROC performances will be higher than the outcome of the regular Logistic regression.

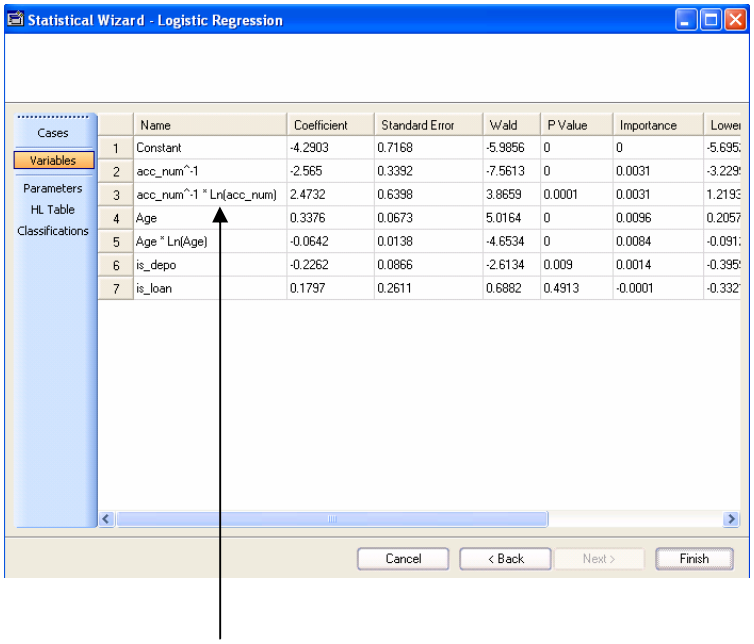


Logistic Regression (F.P.) Analysis Wizard Step 3: Displaying Regression Quick Results

Cases Summary Overview Screen

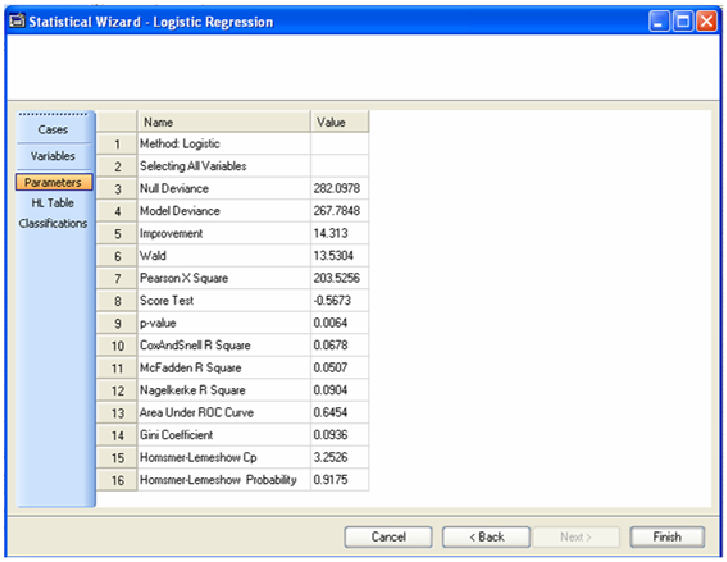


Variables Summary Overview Screen

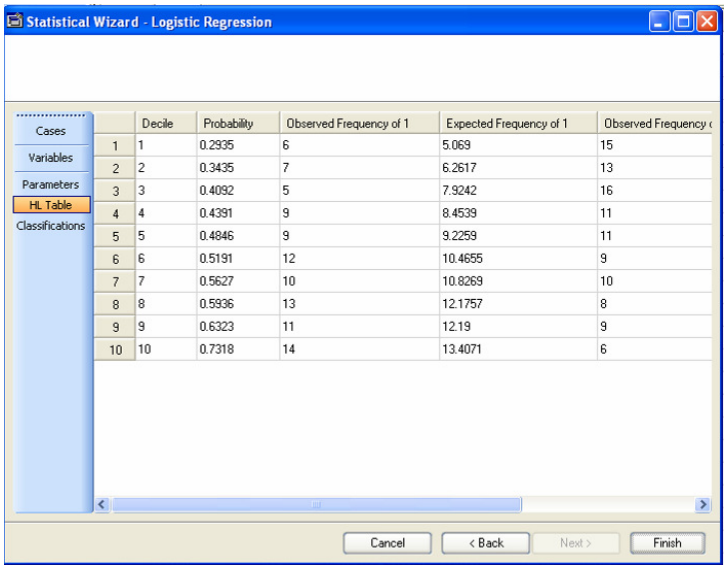


An example of polynomial transformation to the acc_num variable: acc_num^-1 * Ln(acc_num)

Logistic Regression Output Summary Screen

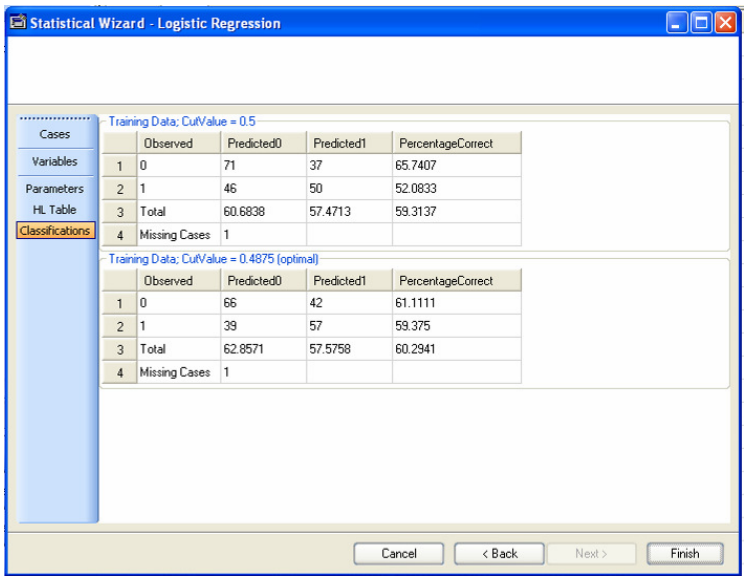


Logistic regression (F.P.) Hosmer & Lemeshow Table Screen



	Decile	Probability	Observed Frequency of 1	Expected Frequency of 1	Observed Frequency
1	1	0.2935	6	5.069	15
2	2	0.3435	7	6.2617	13
3	3	0.4092	5	7.9242	16
4	4	0.4391	9	8.4539	11
5	5	0.4846	9	9.2259	11
6	6	0.5191	12	10.4655	9
7	7	0.5627	10	10.8269	10
8	8	0.5936	13	12.1757	8
9	9	0.6323	11	12.19	9
10	10	0.7318	14	13.4071	6

Logistic regression (F.P.) Classification Optimization Screen



Training Data; CutValue = 0.5

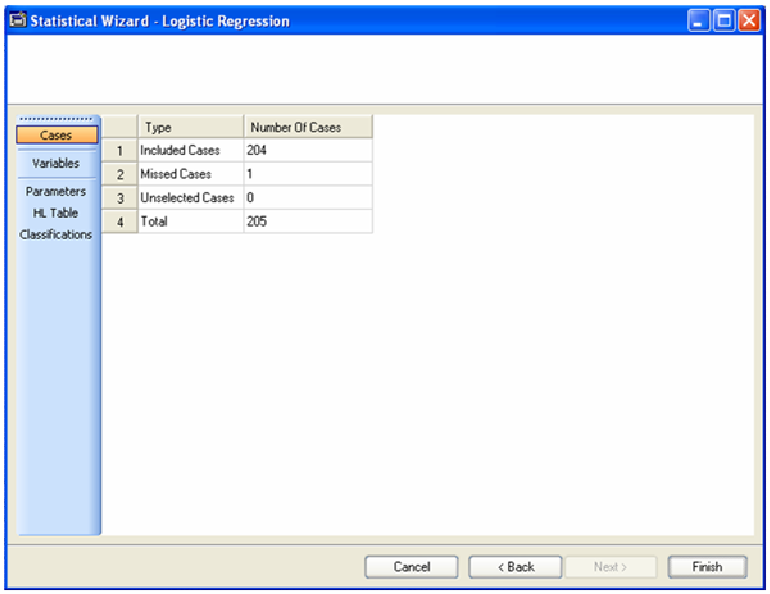
	Observed	Predicted0	Predicted1	PercentageCorrect
1	0	71	37	65.7407
2	1	46	50	52.0833
3	Total	60.6838	57.4713	59.3137
4	Missing Cases	1		

Training Data; CutValue = 0.4875 (optimal)

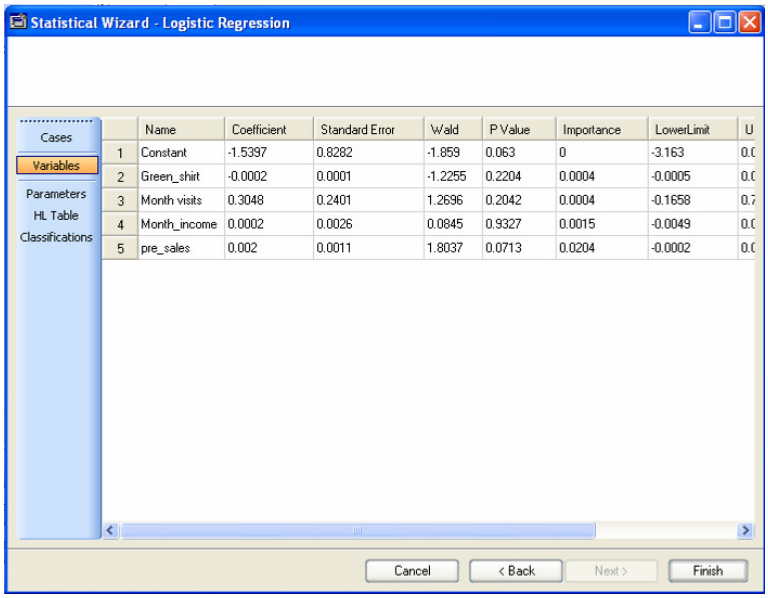
	Observed	Predicted0	Predicted1	PercentageCorrect
1	0	66	42	61.1111
2	1	39	57	59.375
3	Total	62.8571	57.5758	60.2941
4	Missing Cases	1		

Logistic Regression (F.P.) Analysis Wizard Step 4: Displaying Regression Quick Results

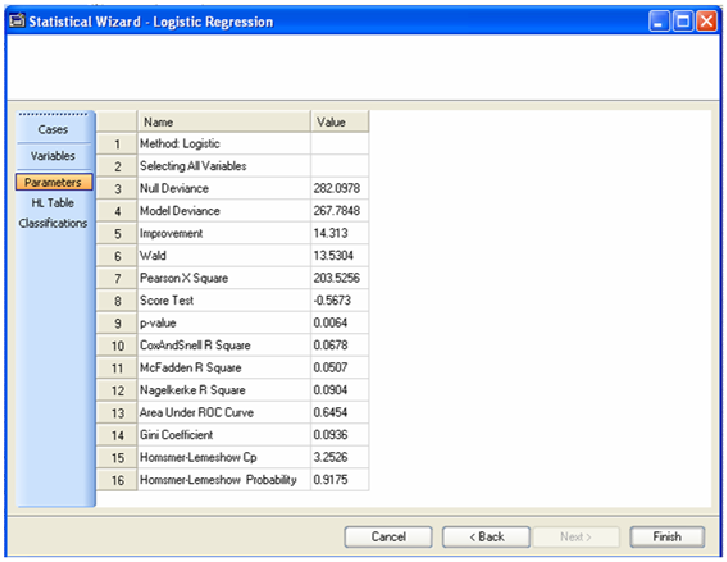
Cases Summary Overview Screen



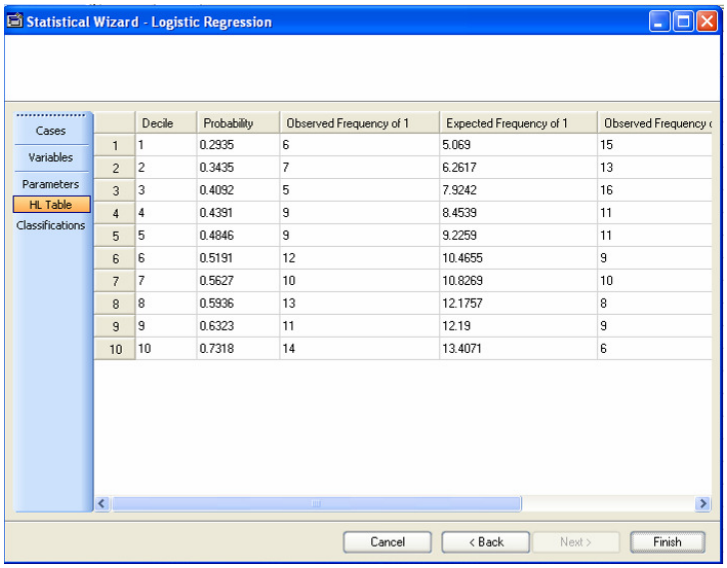
Variables Summary Overview Screen


Statistical Wizard - Logistic Regression
Cases
Variables
Parameters
HL Table
Classifications
Name Coefficient Standard Error Wald P Value Importance LowerLimit U
1 Constant -1.5397 0.8282 -1.859 0.063 0 -3.163 0.0
2 Green_shirt -0.0002 0.0001 -1.2255 0.2204 0.0004 -0.0005 0.0
3 Month visits 0.3048 0.2401 1.2696 0.2042 0.0004 -0.1658 0.7
4 Month_income 0.0002 0.0026 0.0845 0.9327 0.0015 -0.0049 0.0
5 pre_sales 0.002 0.0011 1.8037 0.0713 0.0204 -0.0002 0.0
Cancel
< Back
Next >
Finish

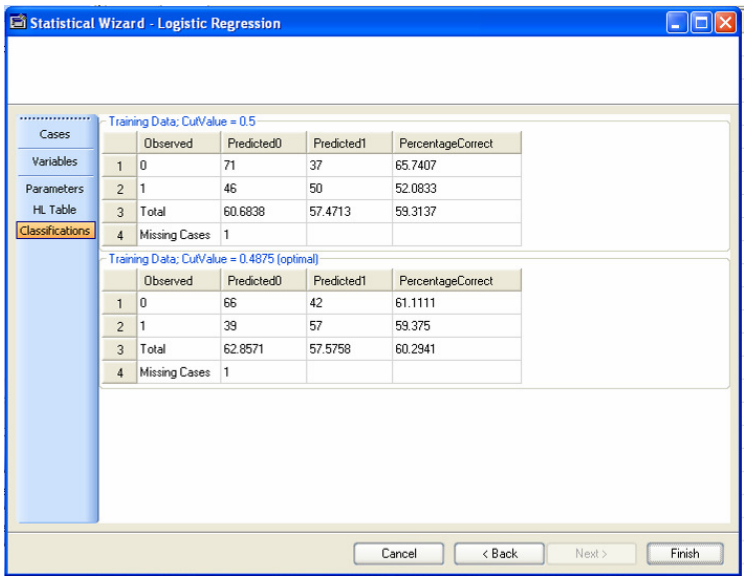
Logistic Regression (F.P.) Output Summary Screen



Logistic Regression (F.P.) Hosmer & Lemeshow Table Screen



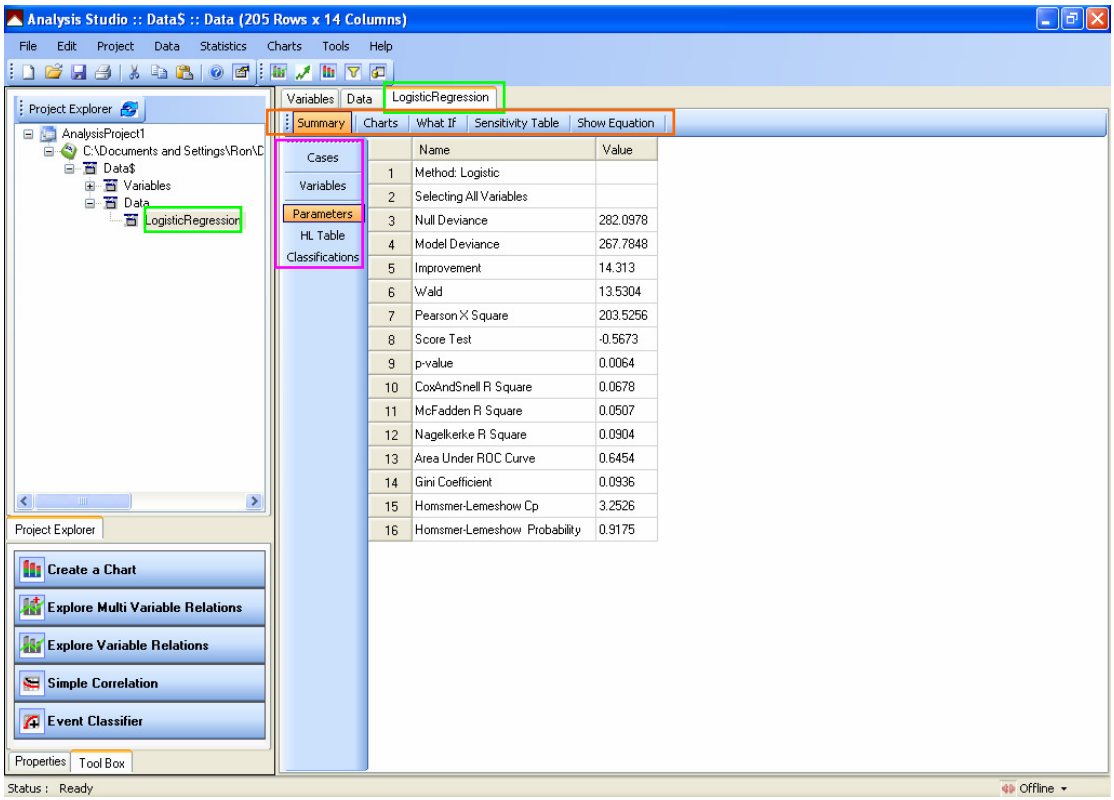
Logistic Regression (F.P.) Classification Optimization Screen



Working with Logistic Regression (F.P.)

After finishing the Logistic Regression (F.P.) Wizard, the software framework is ready to allow the user to inquire and calculate further regressions based on the Wizard results.

Logistic Regression (F.P.) Framework View



The Logistic Regression (F.P.) has the same framework as the regular Logistic regression framework and it contains five main frames. Each of them has its own sub options (marked by a pink frame for the purpose of this explanation).

Logistic Regression (F.P.) Framework Options:

1. Summary
    a. Cases
    b. Variables
    c. Parameters
    d. HL Table
    e. Classification
2. Charts
    a. ROC
    b. Cut points
    c. Gain
    d. Lift
    e. X Diagnostics
    f. Y Diagnostics

g. Cases
h. Hits/Misses
i. Hits Ratio
j. Misses Ratio

3. What-if
    c. Multiple variables What-if calculator
4. Sensitivity
    d. Multiple variables Sensitivity calculator

4.1.5.12 Cox Regression

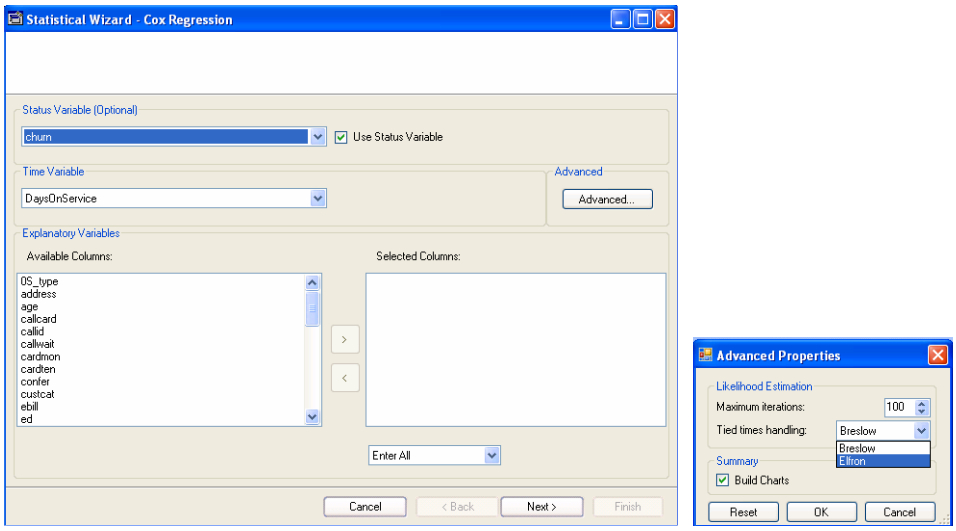
The Cox regression is a time-to-event modeling method. It uses predictor variables to compute a regression mode. For example, a researcher can construct a model showing the length of time for a web site membership in relation to its member's operating system, membership time span or profession category.

The software permits two main model contractions, the first model estimates the survival success based on one or more variables, and the second model estimates the survival success based on one or more variables with a *Status variable* that is the event the researcher would like to analyze.

An example of first model contraction is: what is the chance that a 45 year old customer who lives in a high socio-economic neighborhood has a 50-day active membership? In this model the user should leave the Status Variable.

An example of second model construction is: do gamers and non gamers have different risks of churning a web site based on number of pages viewed? By constructing a Cox Regression model with the number of pages viewed (per day) and gamer or non-gamer identifiers entered as covariates, the researcher can test hypotheses regarding the effects of being a gamer and number of pages viewed in relation to onset of churning the web site.

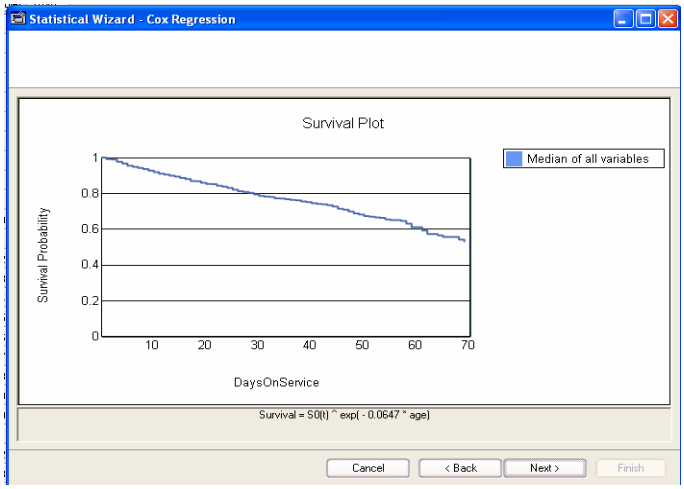
Analyzing Survival Wizard Step 1: Selecting the Appropriate Model Parameters



The first Wizard screen contains two main frames, three combo boxes and an *Advance* settings button. The top combo box titled *Status Variable* is a binary target variable of the analysis process. It can be ignored if the user wishes to explore data regarding the time variable without taking into account a status variable. The *Time Variable* is the variable that contains the number of periods that the user wishes to analyze (i.e. Number of months belonging to the service, Years with company etc.).

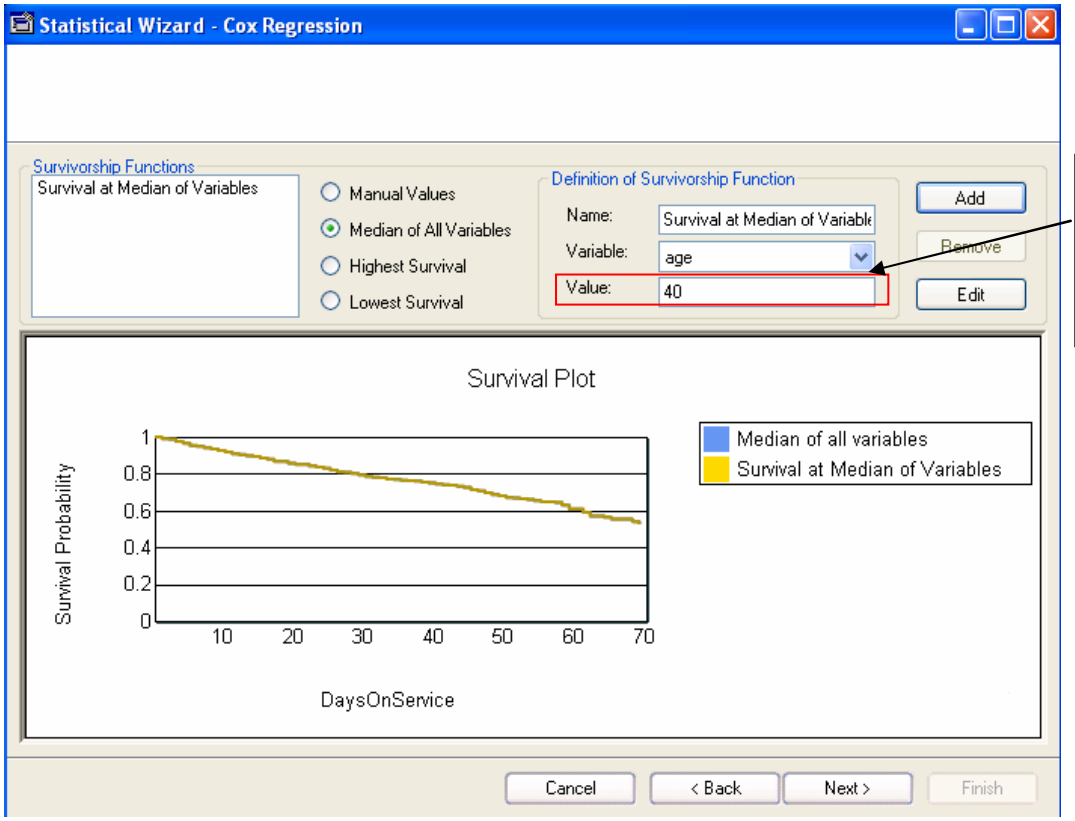
The *Explanatory Variables* section contains two frames: *Available Columns* frame that contains all numeric variables of the data set and a *Selected Columns* frame that contains the variables that the user has selected for the *Available Columns* frame. For well trained users there is an *Advanced* button which allows the user to select methods of treating time-constrained events.

Analyzing Survival Wizard Step 2: Viewing Main Survival Chart



In the above screenshot the software's Cox Wizard displays the Survival Probability as a function of the *Time Variable* that is being affected by the explanatory variable/s.

Analyzing Survival Wizard Step 3: Survival Fine-tuning and Manual Settings



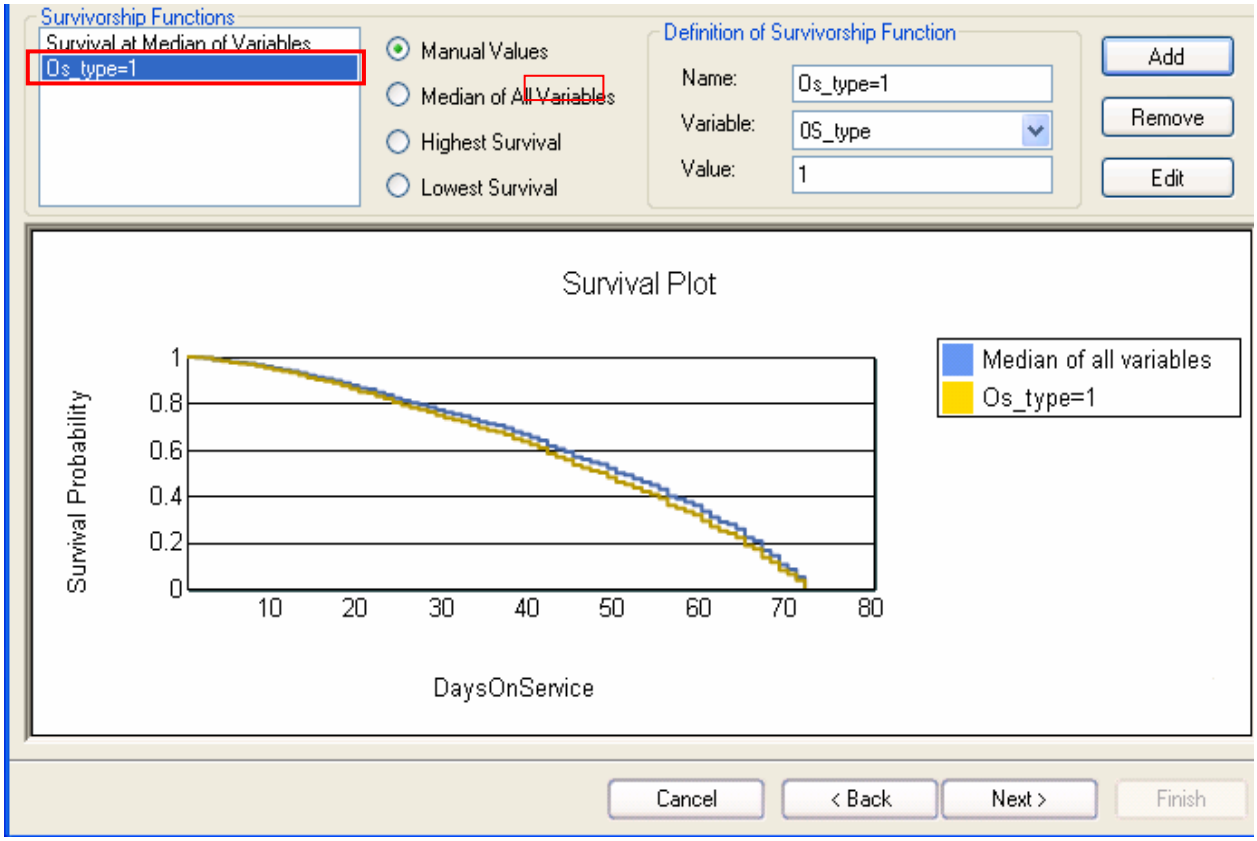
The user can manually set the value of the variable

By default, the Wizard screen above displays the survival function as a function of the variable's Median value (in this example the median age for this data set is 40). On occasions, the analyst would like to inspect other values of the explanatory variable; the software allows **unlimited selection of values for each variable** that is included in the Survival modeling process.

For each Survival analysis the user can add a new survival calculation that is based on the Survival function as produced in the first Wizard screen.

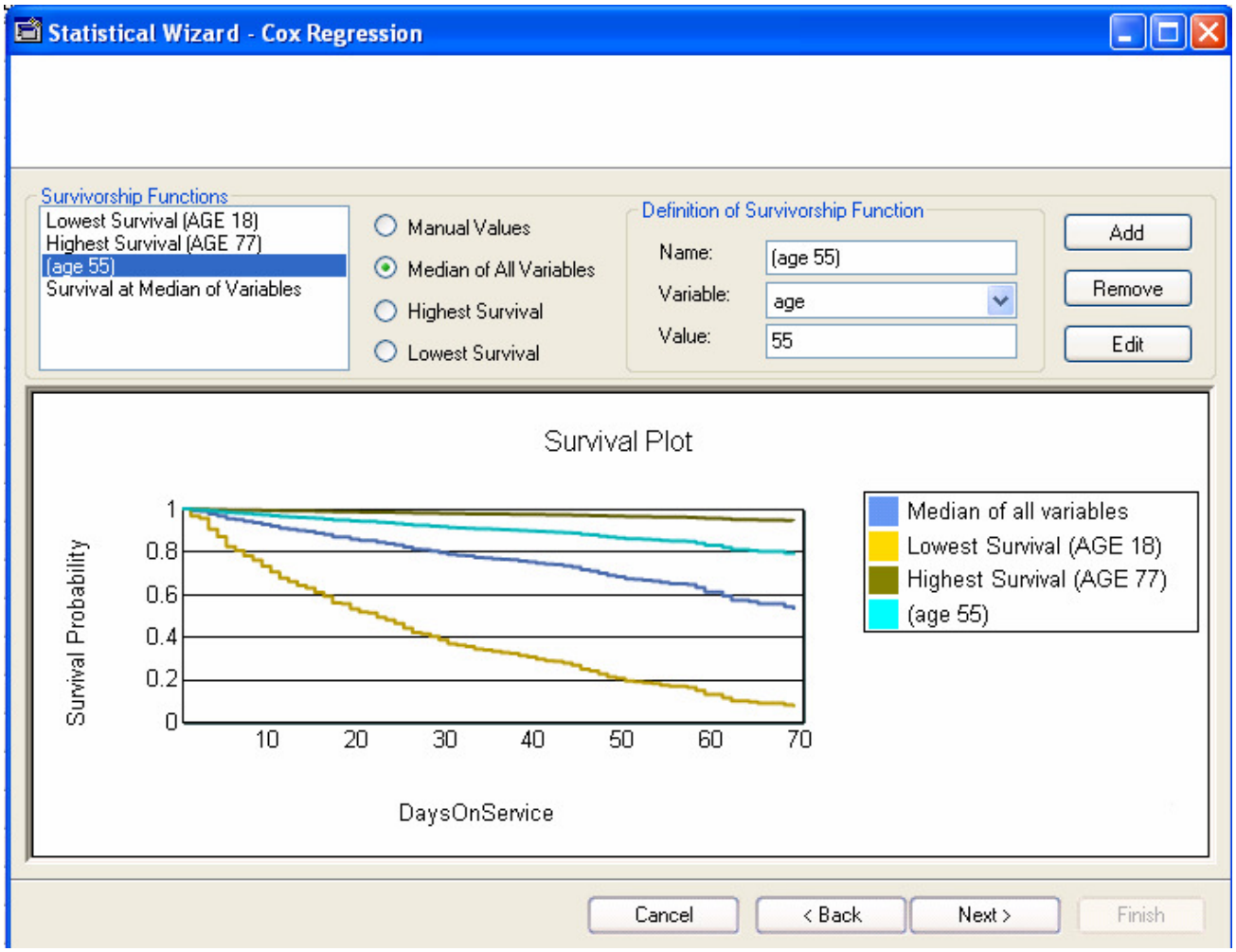
In order to add variable values to the survival variable the user should select the *Manual Values* option and enter a new value.

In order to change the function name or variable value of a survival function the user should select the desired survivorship function from the *survivorship functions* frame and click the *Edit* button.

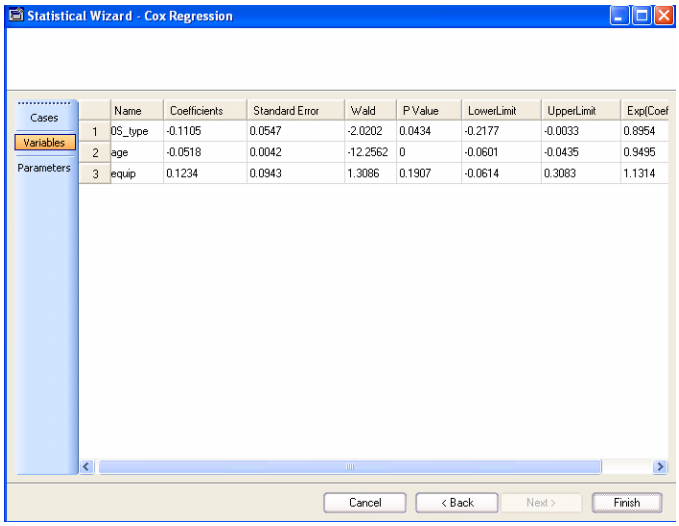

Survivorship Functions
Survival at Median of Variables
Os_type=1
Manual Values
Median of All Variables
Highest Survival
Lowest Survival
Definition of Survivorship Function
Name:
Os_type=1
Variable:
OS_type
Value:
1
Add
Remove
Edit
Survival Plot
Survival Probability
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
DaysOnService
Median of all variables
Os_type=1
Cancel
< Back
Next >
Finish

For example: in the screen below the researcher chose to build four Survival functions for different age values. In this case the Wizard plots three additional functions and the researcher has written down the preferred names for them.

Customizing Survivorship Functions:



Analyzing Survival Wizard Step 4: Viewing Quick Results



The Wizard screenshot above displays the predicators of the Survival function. The software automatically calculates survival probability in the *Survival Model* tab while the Survival modeling Wizard displays the statistical parameters for each variable:

- Name
- Coefficient
- Standard error
- Wald
- P value
- Lower limit
- Upper limit
- Coefficient's Exponent: *Exp (Coefficient)*

Here is an interpretation of the variables Coefficient's Exponent:
The meaning of having *Exp (Coefficient)* 0.9495 for the Age variable under a Churn *Status Variable* is that every additional year the Churn probability is reduced by 100%-(100%*0.9495) = 5% .

The meaning of having *Exp (Coefficient)* 2.5821 for the *Equip* categorical variable under a Churn *Status Variable* is that having special equipment increases the Churn probability by 258.21%.

Survival Tab

The *Survival Tab* contains five sub frames that allow the user to view statistical results as well as conduct Survival calculations based on the Survival Wizard outcome.

The five sub frames are:

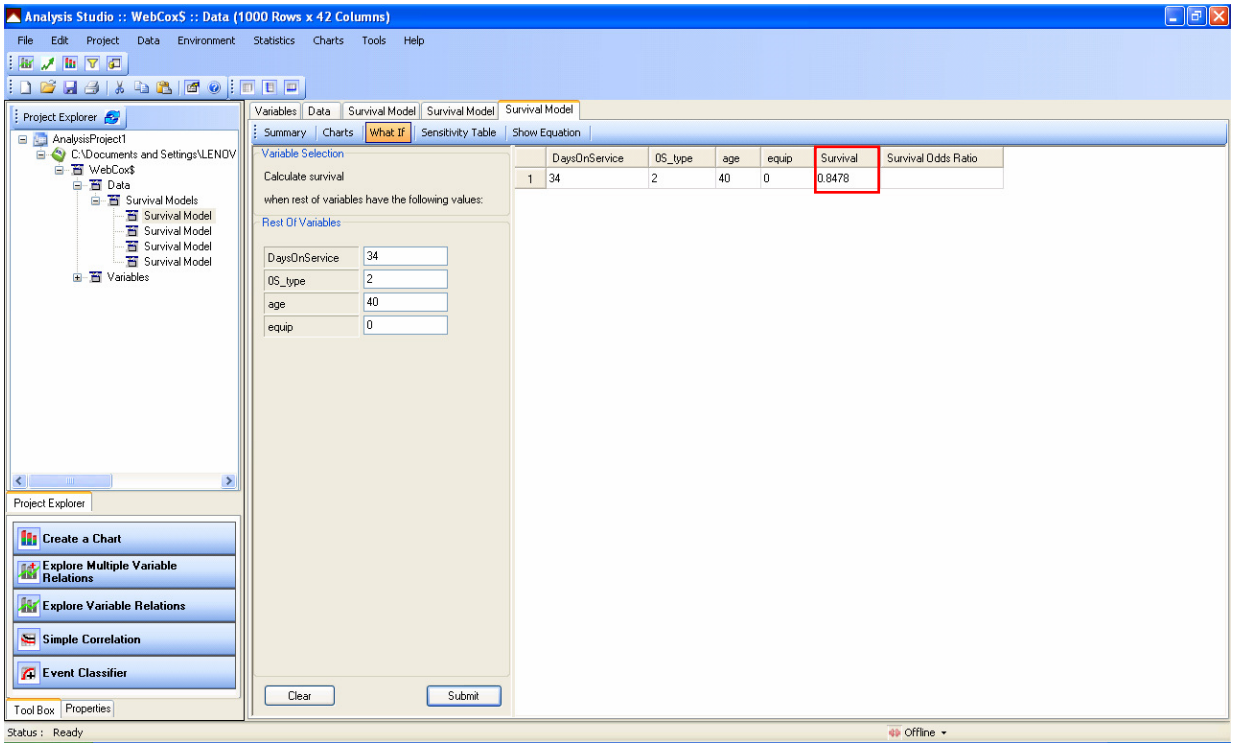
1. Summary
    a. Cases
    b. Variables
    c. Parameters
2. Charts
    a. Survival
    b. Hazard
3. What-if analysis
4. Sensitivity analysis
5. Show equation

An example for What-if analysis and Sensitivity Table:

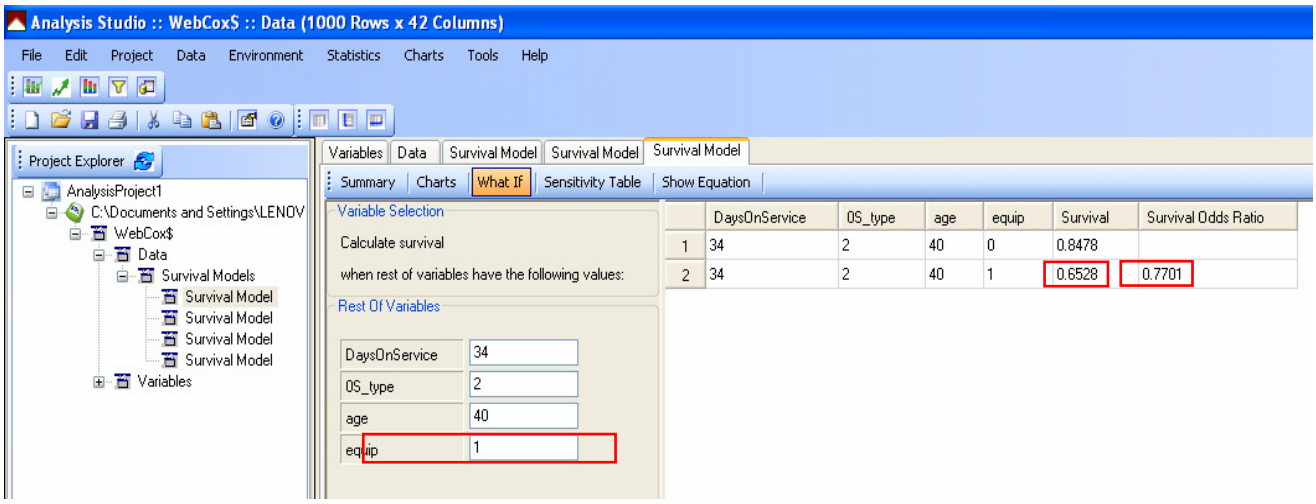
In this example the software has calculated a survival function to a Churn *Status Variable* Using four variables:

1. DaysOnService: number of days that the customer is with the company
2. Os_type: Operating system RAM power
3. Age: age of customer
4. Equip: a categorical variable indicating presence of special computer equipment ("1" means that there is special equipment)

The default values for the variables are their median values. By clicking the *Submit* button the software calculates the Survival success (non-churn) probability. In this case a customer who is subscribed to the service 34 days, has a medium RAM powered OS type, is 40 years old and doesn't have any special equipment, has 84.78% chance to stick with the company.



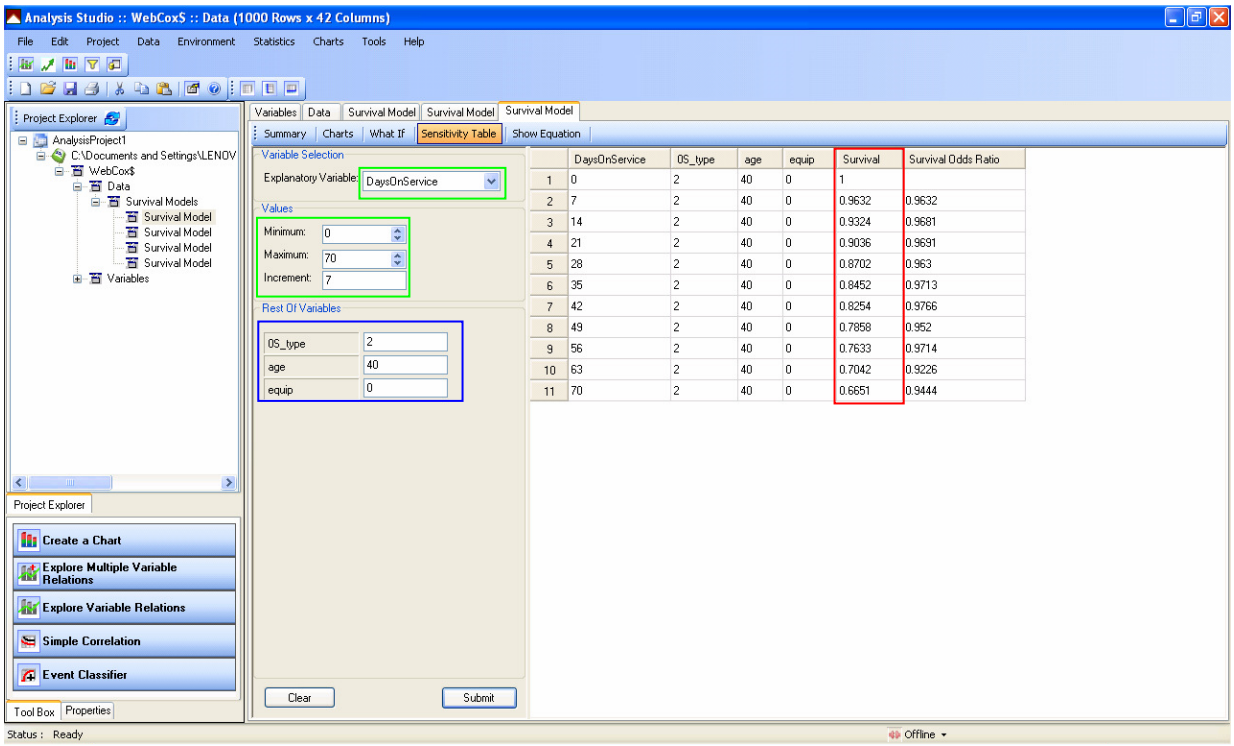
Let us look at the chances of a customer that has special equipment:



That is a steep drop in the chance to stay with the company: A professional customer (who has special equipment) has only 65.28% chances to stick with the company with the given parameters.

Using the Sensitivity Table:

This example has the same basic survival parameters and variables that the *What-If* example has. The difference between them is that the *What-If* analysis is designed to analyze specific variable values and the *Sensitivity Table* is designed to view how the target calculation changes according to one single variable value change.



The screenshot above displays the Survival rate (red frame) as it changes with having the Day of Service variable's values change (green frame). The rest of the variable values remain constant (blue frame).

4.1.5.13 Forecasting and Time Series

The Forecasting and Time Series module has six forecasting models:

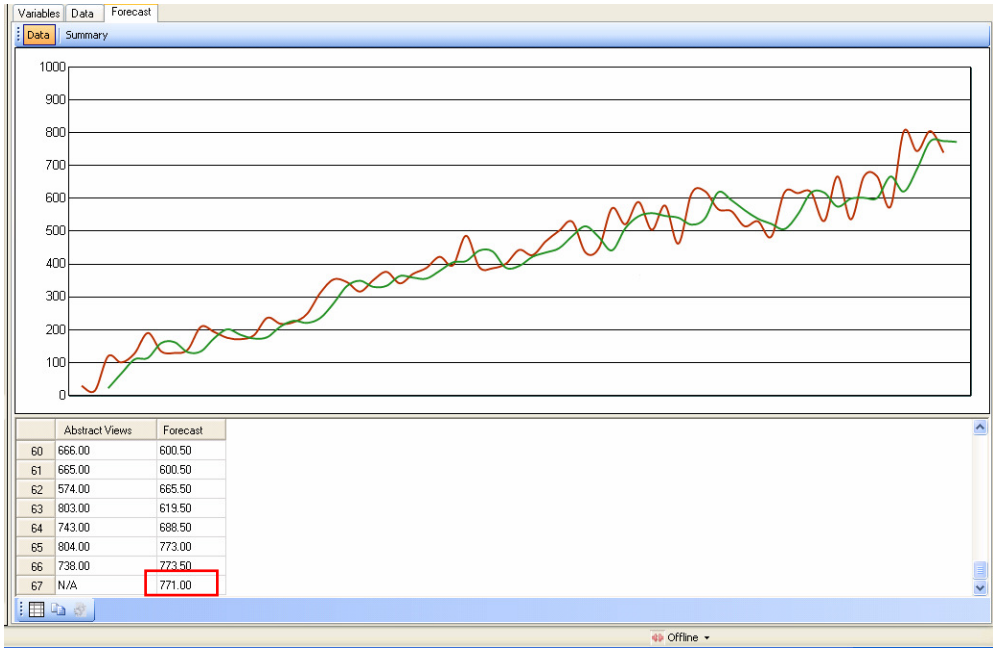
Random Series: the software selects a value randomly from data equally distributed (suitable mainly for predicting stocks and currency rates). The user can only select the number of time periods that the software will calculate the appropriate forecasting values for them. It is advised that in this model the *Num of Periods* will be set to 1.

Forecasting Optimization: The software includes several optimizing mechanisms that reduce the forecasting errors (*MAE, RMSE, MAPE*) as well as automatic optimization mechanisms for factor parameters used by four out of six forecasting models. By using those mechanisms the user can save a lot of time and effort selecting the best factor value.

Treating Seasonality and Trends: there are four Models (*Moving Average, Exponential Smoothing, Holt's Model, Winter's Model*) that contain a seasonal factoring mechanism that the user can use when seasonality effects are suspected. If the software is unable to detect one of the seasonality pattern types a message box will appear with the message: "The requested seasonality pattern was not found." By default all models are in non-seasonal mode enabling seasonality to be controlled by the user from one of the *Seasonality Type* frame.

An optimized mechanism for best factors selection is also included for the four Models.
Treating Trend: Two Models (*Holt's model* and *Winter's Model*) have a *Trend factor* option that can be optimized automatically or manually.

Displaying the *Forecasting* Results and Predicted Values: the F*orecasting* tab includes a *Data* tab and a *Summary* forecast tab. The data shown in the Data tab include the analyzed column and the predicted value or values in case the *Num of Periods* is set to more then one.



The red-framed number in the above screenshot is the predicted value that was calculated according to the forecasting model and the parameters selected by the user in the forecasting Wizard step 1.The *Summary* tab contains the main performance results of the model.

There are three model performance parameters that are included in the *Summary* tab:
MAE: The mean absolute error (MAE) function is a weighted average of the absolute errors between the actual and the predicted values. The software was designed to reduce its function to the minimum using an automatic mechanism.

RMSE: The root mean absolute error (MAE) function is a weighted average of the absolute root errors between the actual and the predicted values. The software was designed to reduce its function to the minimum using an automatic mechanism.

MAPE: The mean absolute error does not depend on the units of the forecasted data column but is always stated as a percentage. This makes model comparison an easy task: for example a model that has 8.5 MAPE is better then 10 MAPE because in the first model forecast is off on average by 8.5 % and the second model is off on average by 10%.

Random Walk: the software selects a value not randomly but by the equally distributed stages between their data and their successors (suitable mainly for predicting stocks and currency rates). The user can only select the number of time periods that the software will calculate the appropriate forecasting values for them. It is advised that in this model the *Num of Periods* will be set to 1.

Moving Average: the software calculates the average of the values during the time frame the user has selected. The user can change two parameters: *Num of Periods* which is the number of periods that the user wants to predict and *Span* value which is the number of time periods that are considered for the forecasting. This means that the lower the span is the closer the forecast is to recent time periods. For example a span of 2 for a monthly data means that the average calculations will use the last two months' values.

Exponential Smoothing: In this method the software calculates smoothing data series and seasonality. Recent observations are given relatively more weight in the forecasting calculation than older observations and vice versa. The software includes a factors optimizer in order to find the trend factor that minimized the four errors functions.

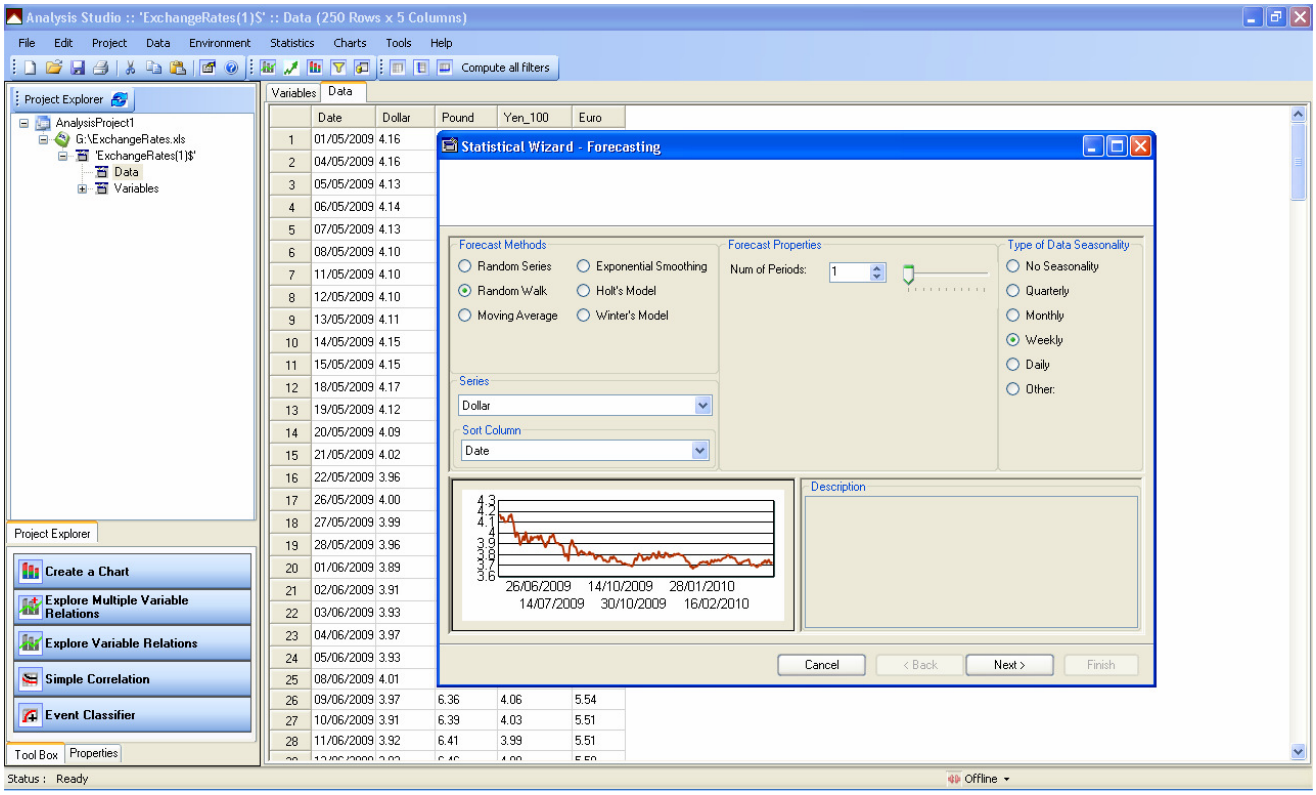
Holt's model: In this method the software calculates the natural data trends, smoothing data series and seasonality. The weight of the trend depends on the *trend factor* that can be determent manually or automatically.

Winter's Model: this method is based on *Holt's model* and was designed to handle a seasonally-patterned time series in a more accurate manner. To enable the model the user should select one of the seasonality types. This model uses 3 smoothing constants:

- One for the signal,
- One for the trend and
- One for seasonal factors.

The user can choose three factor levels manually or select “Optimize Factor” for automatic optimization. The three factors are: *Smoothing Factor*, *Trend factor*, and *Seasonality factor*.

Forecasting Wizard Step 1: Selecting the Appropriate Model Parameters and Data



There are four parameters that the user should select in order to generate the forecasting:

First step: select the desired forecasting model.

Second step: select the target variable (i.e. Dollar/Pound rate) and the “Sort Column” that the model will be built upon.

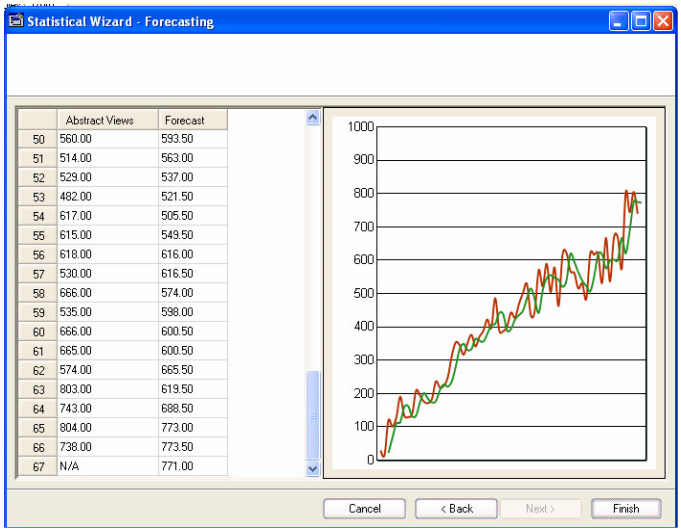
Third step: select the forecast properties which include the number of periods to be calculated and the other unique properties of the selected model.

Fourth step: select type of seasonality, if any.

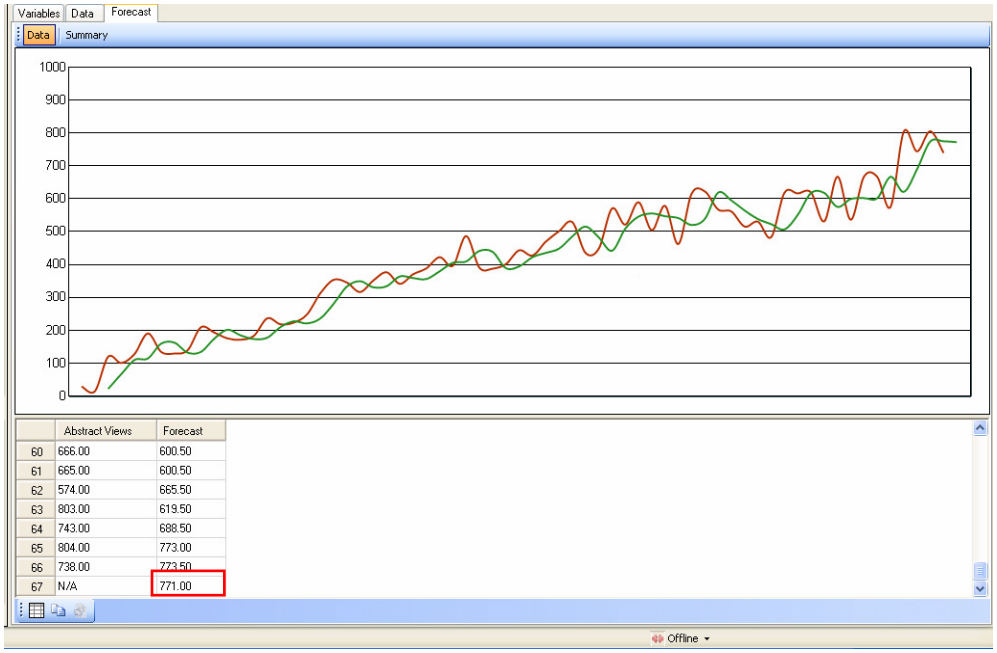
Fifth step: select from the “Series” combo box the desired column. The Series combo box is for selecting the desired column from the data set. A part of the column’s values will be shown in the chart frame for quick inspection.

In the above screenshot the time series the user would like to forecast is Dollar/NIS rate and the sort column is the Date column that contains date on a daily basis.

Forecasting Wizard Step 2: Inspecting the Results



The green line is a predicted line that the software produces in order to display the model's performance even before inspecting mathematical results.



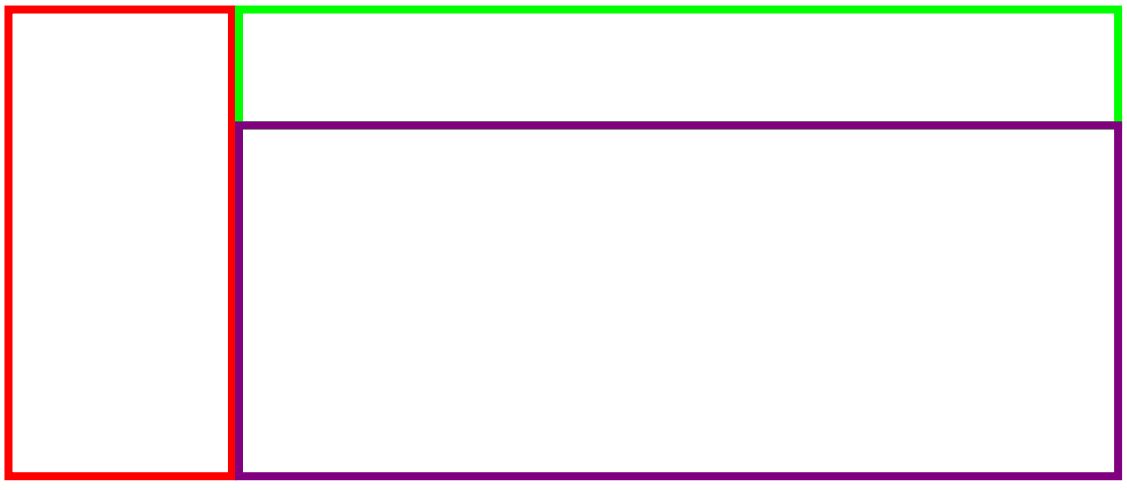
The screenshot above displays the Forecasting tab which includes a Data tab and a Summary forecast tab.

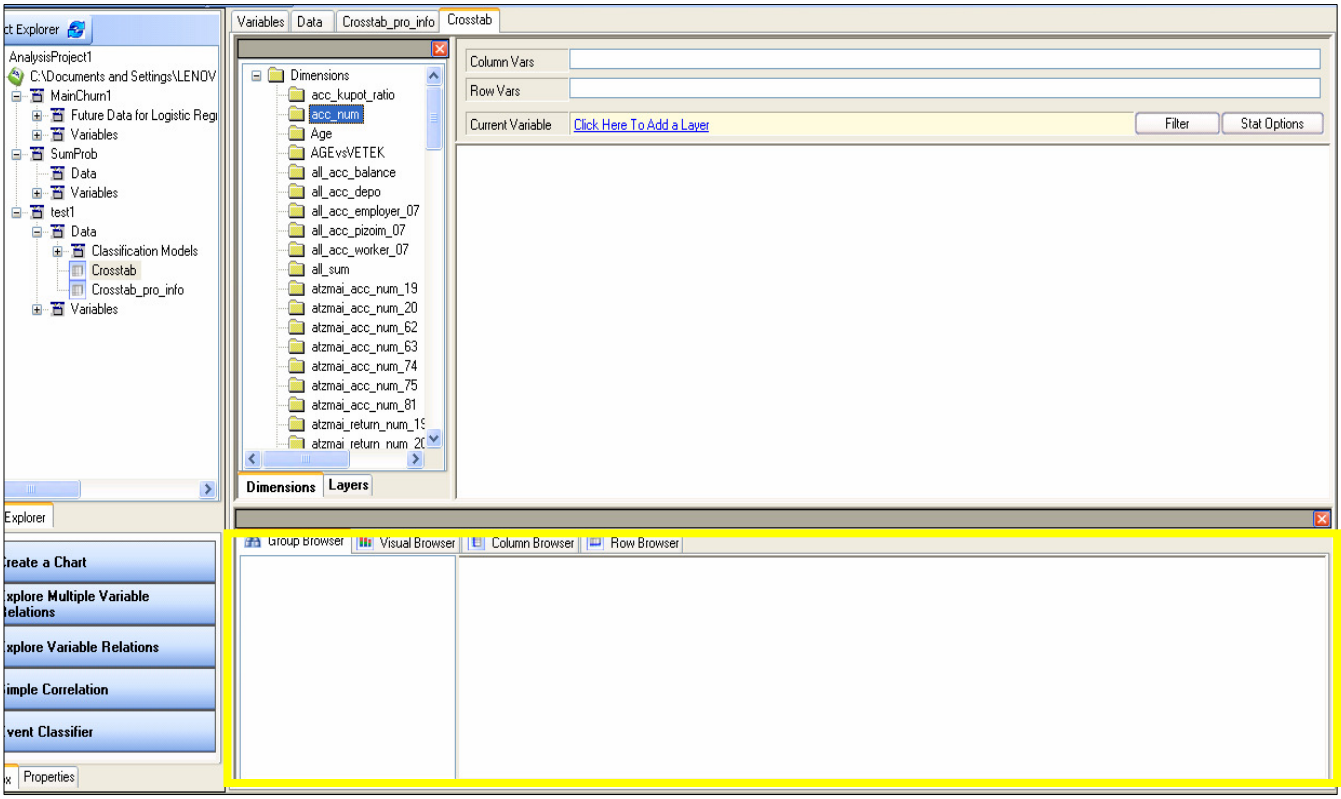
The Data tab data includes the analyzed column and the predicted value or values in case the "Num of Periods" is set to more than one. The red framed number is the predicted value that was calculated according to the forecasting model and the parameters selected by the user in the forecasting Wizard step 1. It is the 67 time period value that is based on 66 historical time period's values. The "Summary" tab contains the main performance results of the model.

4.1.5.14 Cross-Tab Engine

The *Cross-Tab* engine is designed to aggregate data according to specific views that the user defines. The software *Cross-Tab* engine has no view limitations but it is recommended that no more than three layers be used per dimension (Column variables or Row variables).

Building a Cross-Tab View (Step 1):





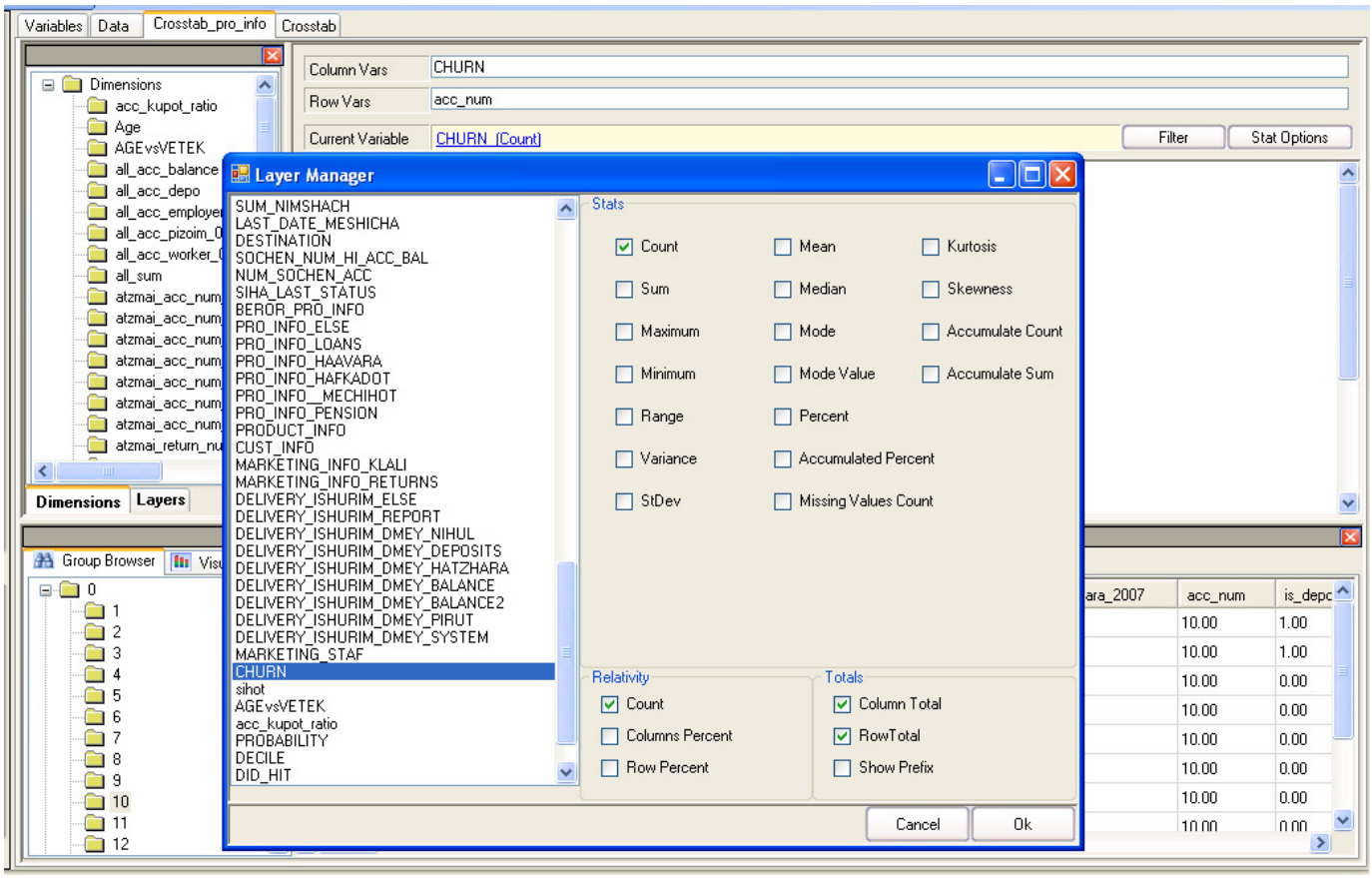
In the screenshot above there are four colored frames for each of the main frames contained by the *Cross-Tab* engine tool.

The red frame is the frame that contains all variables that can be sliced in the main framework (purple frame). The user can drag and drop variables into the green frame that contains five elements:

1. *Column Vars* frame – the user can drag and drop variables into this frame and they will become the header.
2. *Row Vars* – the user can drag and drop variables into this frame and they will be the horizontal rows.
3. *Current Variable* – by clicking this button an operational screen *Layer Manager* is opened and the user can select the desired variable or variables to be displayed on the cross-Tab table. The user can select the desired statistical measure on the same screen. There are 24 statistical measures that can be computed for the selected variables.
4. *Filter* – enables the user to filter out unnecessary values.
5. *Stat Options* – enables the user to select statistical measurements that will be displayed in the *Cross-tab* table view.

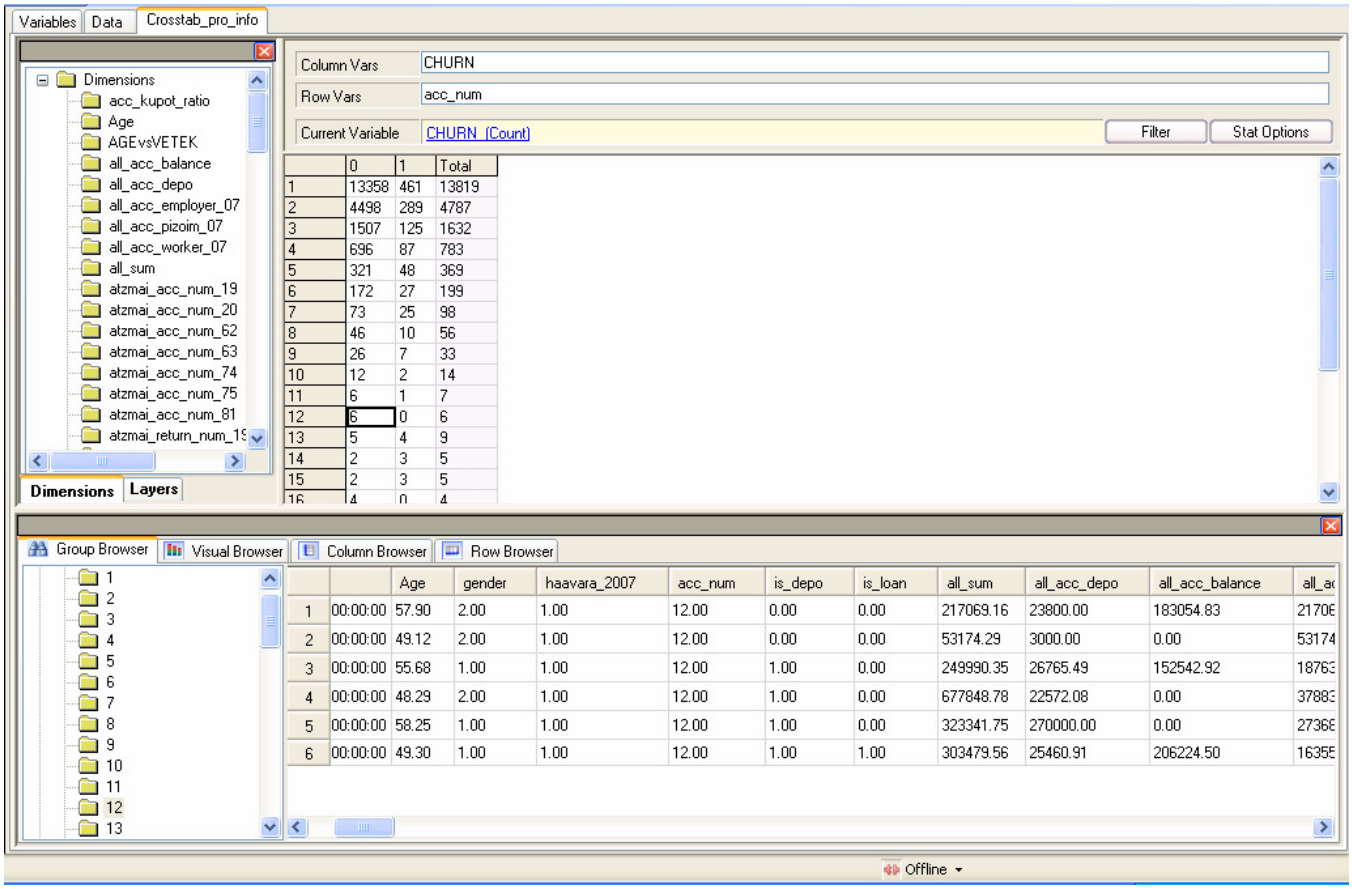
"Cross-tab table is the main viewer displaying the output of the selected dimensions. The Cross-tab view shows statistical measures that the user has selected (marked by the purple frame) and the Browser rows and data display the data beyond the table (marked by the yellow frame).

Building a Cross-Tab View (Step 2):



The screenshot above displays the Step 2 View for building an operation. In this operation the actual selections are made by the user. For example, as shown in the above screenshot the column variable is “CHURN”, the Row variable is “acc_num”, the current variable is “CHURN” and the statistical measure is “count”. The totals that were selected are column and row totals.

Viewing a Cross-Tab Table (Step 1):



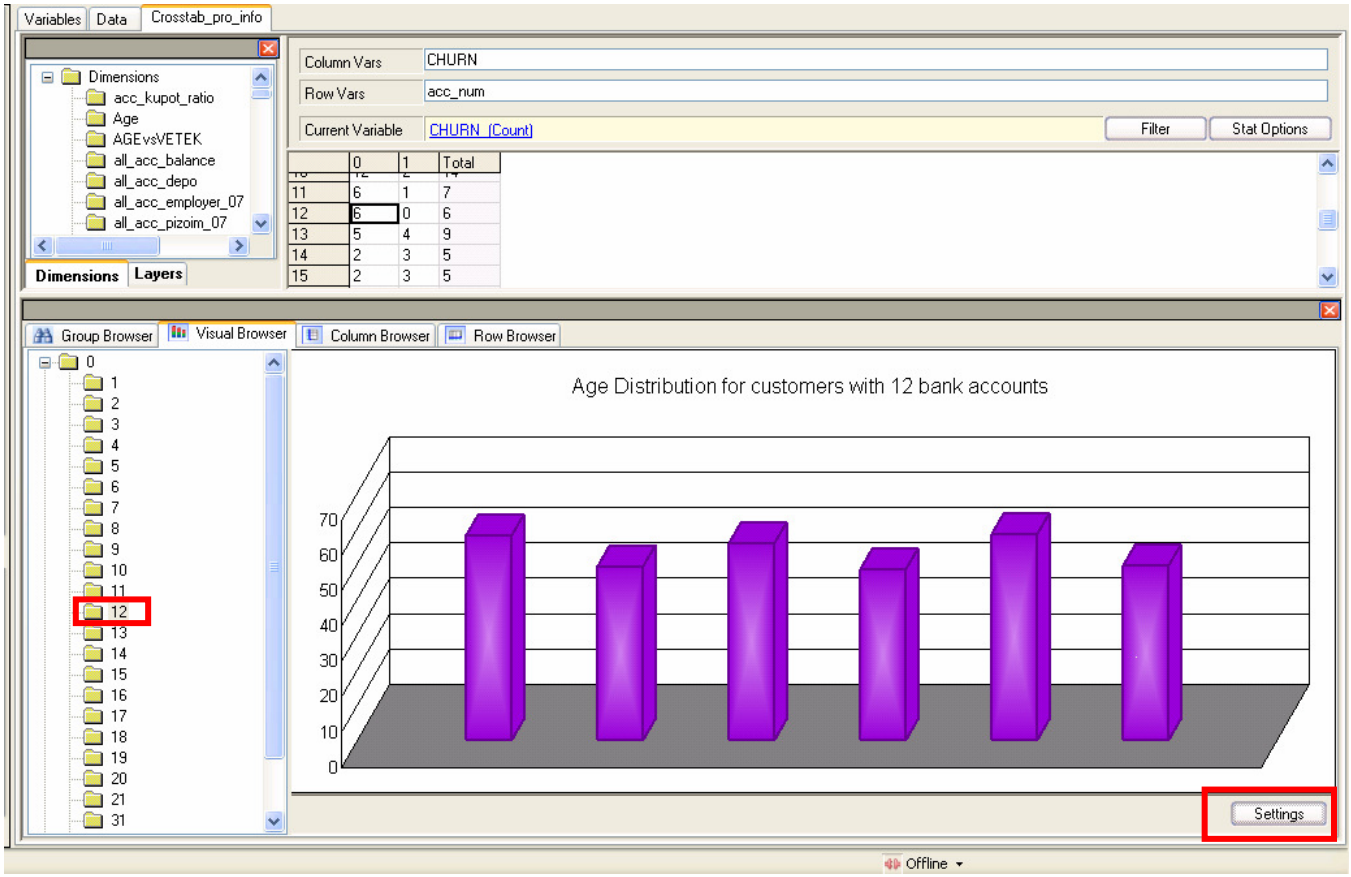
The screenshot above displays a two dimensions view that contains a customer's churn status and the number of bank accounts.

For Example: there are six customers that are not churners and have 12 bank accounts.

Their details are shown in the “Group Browser”. There are four browser options for in-depth inquiry:

- **Group Browser** – displays the rows that are contained in the *Row Vars*. This is an easy way to look for details in the selected group.
- **Visual Browser** – displays a chart of any dimension that the user selects according to the data contained in the selected group. The user can change the chart settings including the dimensions in the *settings* button.
- **Column Browser** and **Row Browser** - each display the data contained in the view according to the column or row headers.

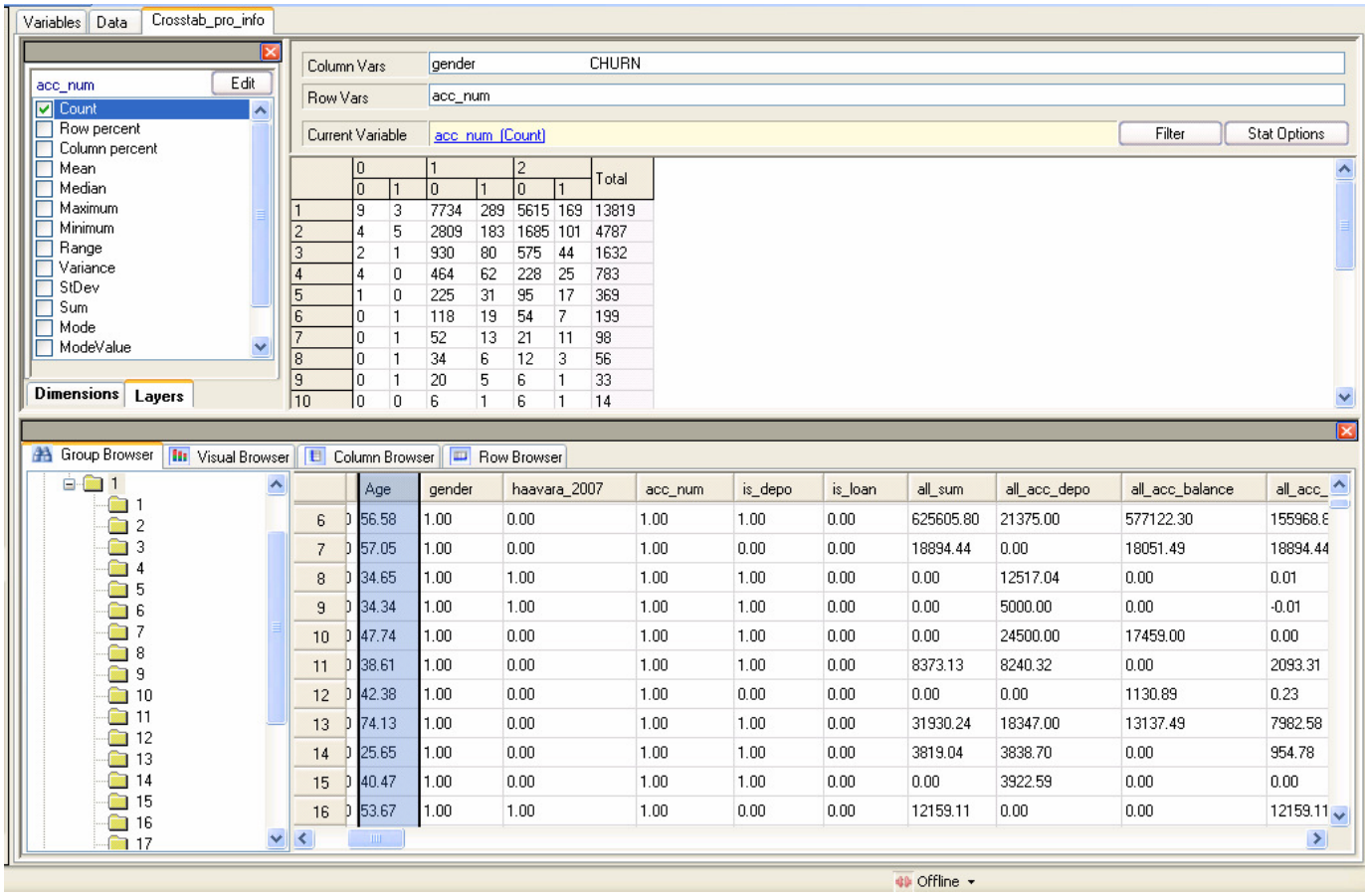
The Use of the Visual Browser:



The screenshot above displays the Age Distribution for customers with 12 bank accounts.

The Chart's header was created using the Chart's settings button. The chart settings cannot be saved and are only for the purpose of immediate inquiry.

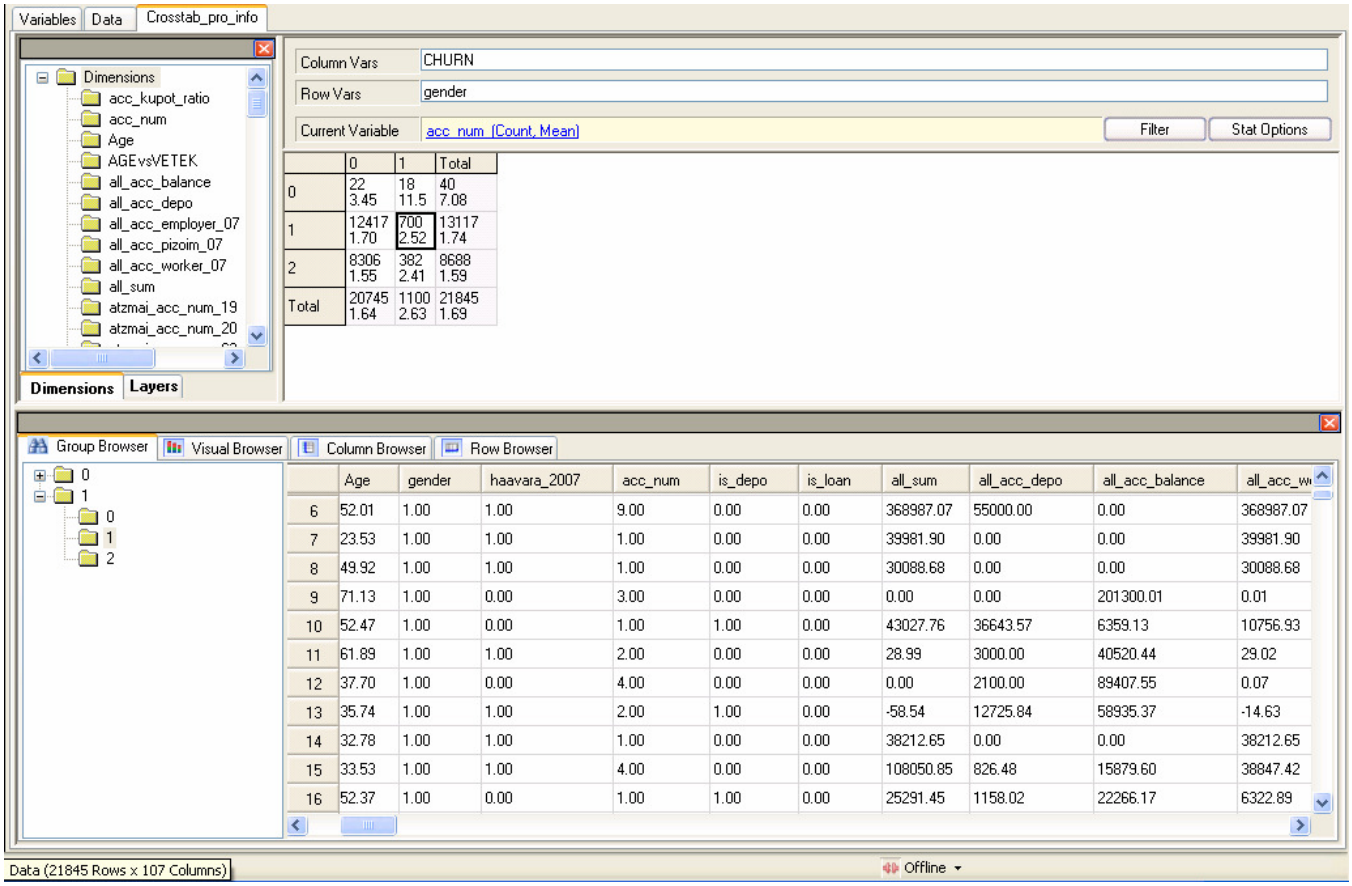
Example for Three Dimensions View



The screenshot above displays a view of bank accounts number with three dimensions:

- The first dimension is The Gender variable that has three categories ("0" for Gender unknown, "1" for males, "2" for females).
- The second dimension has two categories ("0" for non-churners and '1" for churners).
- The third dimension is for a third row dimension (bank accounts number) which slices the six groups. The outcome is a table view that shows all complex data in a simple display.

Example for Two Dimensions View with Statistical Measure for a Different Variable:



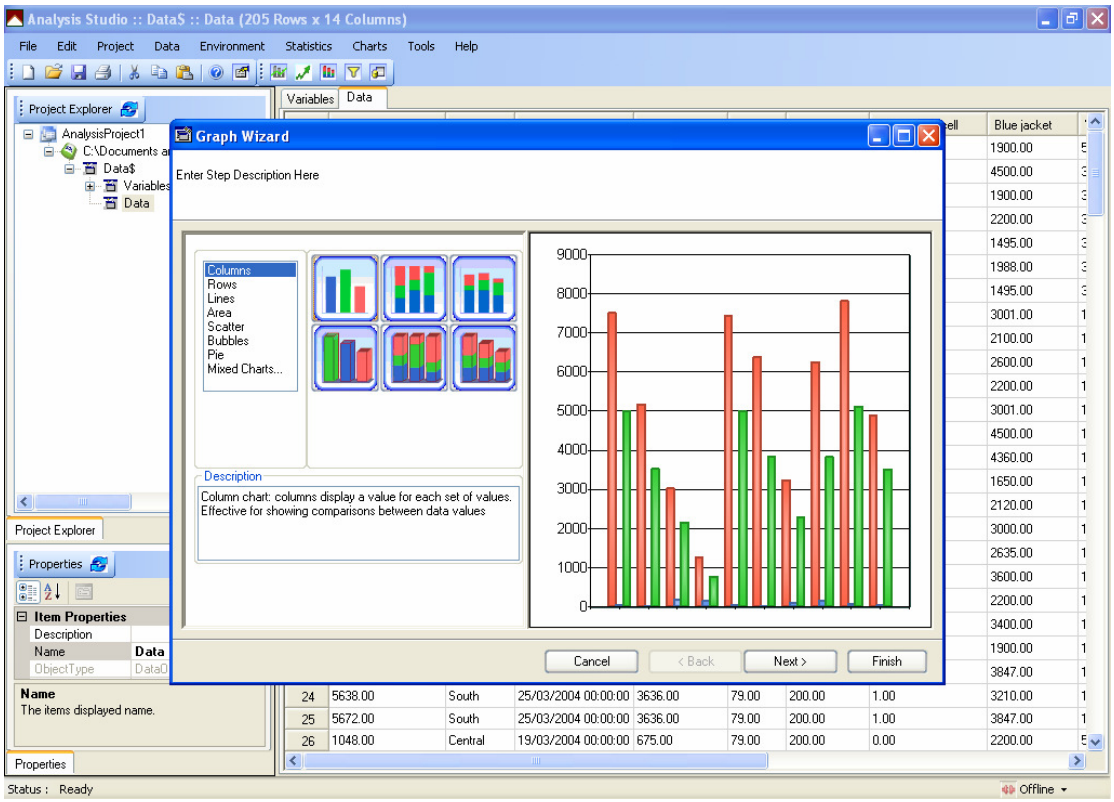
The screenshot above displays a view of bank accounts number and mean over five customers groups: Unknown gender (“0”), males (“1”) and females (“2”) for each group there is a churn status (“0” for non-churner and “1” for churner). As one can notice, the group of male churners has a higher mean of accounts and they are the larger group among all three groups of churners.

4.1.6 CHARTS MENU

4.1.6.1 New Chart

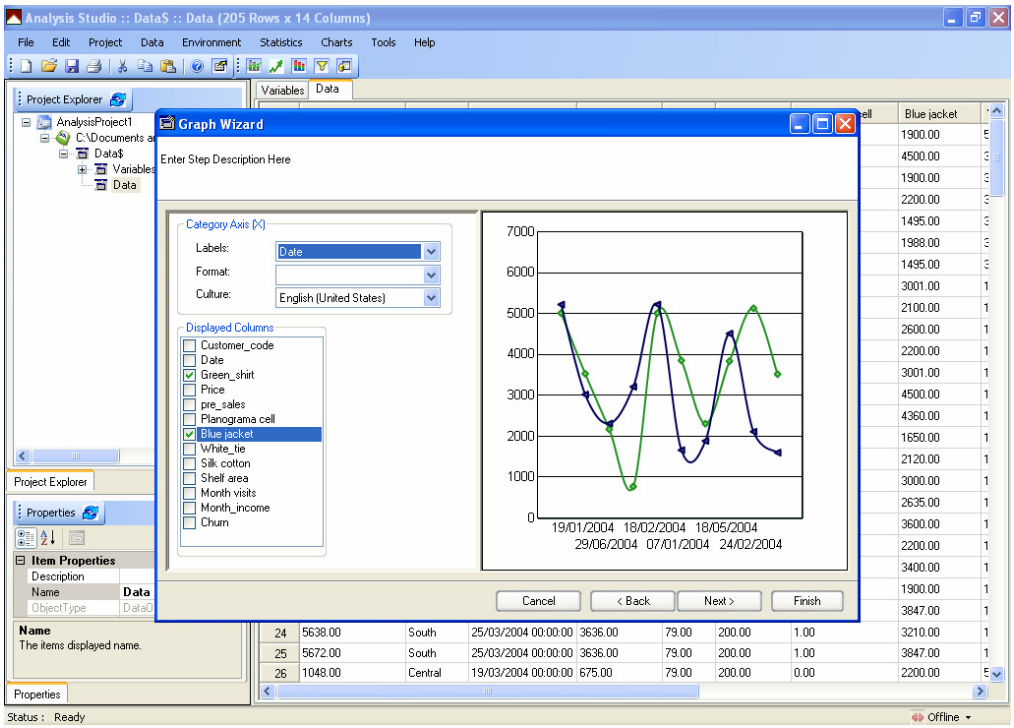
The charts menu contains one entity for creating charts in order to view the data as desired by the user. The user can create charts that are independent of other Analysis Studio procedures.

Charts Wizard Screen 1: Selecting the Desired Chart



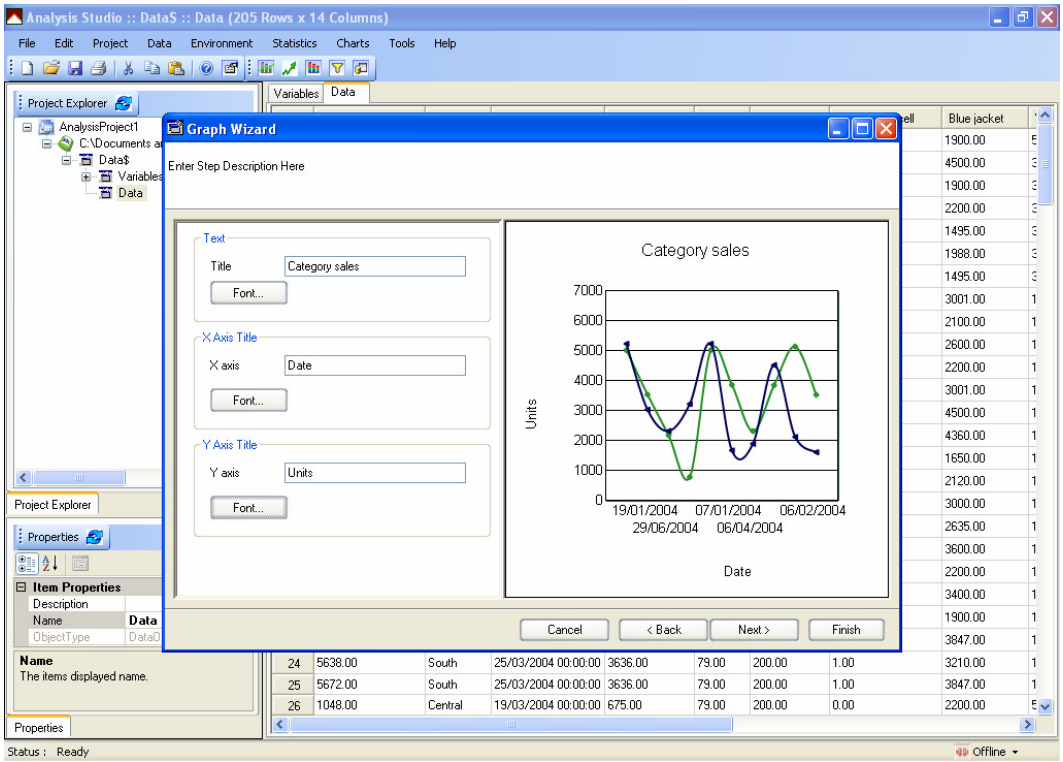
The software contains 6 main chart types and 45 chart sub-types including a mixed charts option. For each chart sub-type there is a description text. By default, the first three numeric variables of the data set are displayed in the right Wizard frame. The data points that are displayed on the right frame are randomly selected; no chart interpretation should be done based on the right frame view only.

Charts Wizard Screen 2: Selecting the Displayed Columns and Axis(X) Properties

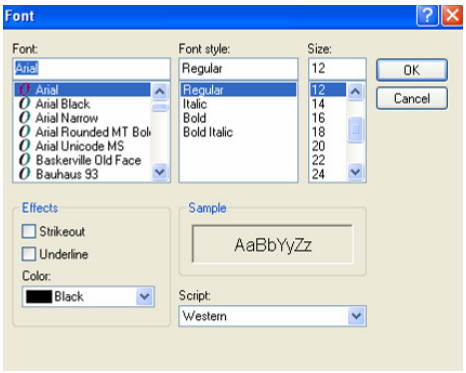


The user can use Category Axis (X) frame to set the X-axis properties as label name, axis format and desired culture. By default, the first three columns of the data set are displayed. However, the user can change columns display to be suited to any other combination.

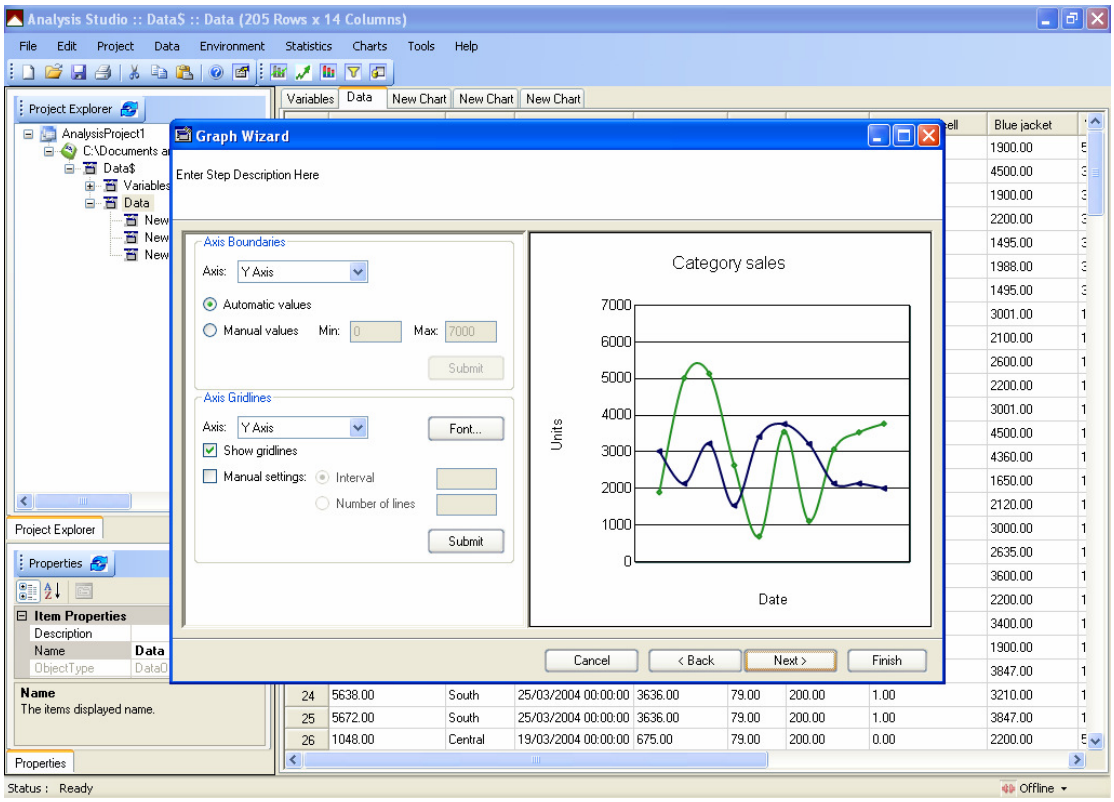
Charts Wizard Screen 3: Adding Titles for the Chart and its Axes



The user can add titles for the chart and the axes; each title can be formatted by using the Font dialog box:



Charts Wizard Screen 4: Changing Axes Boundaries and Gridlines



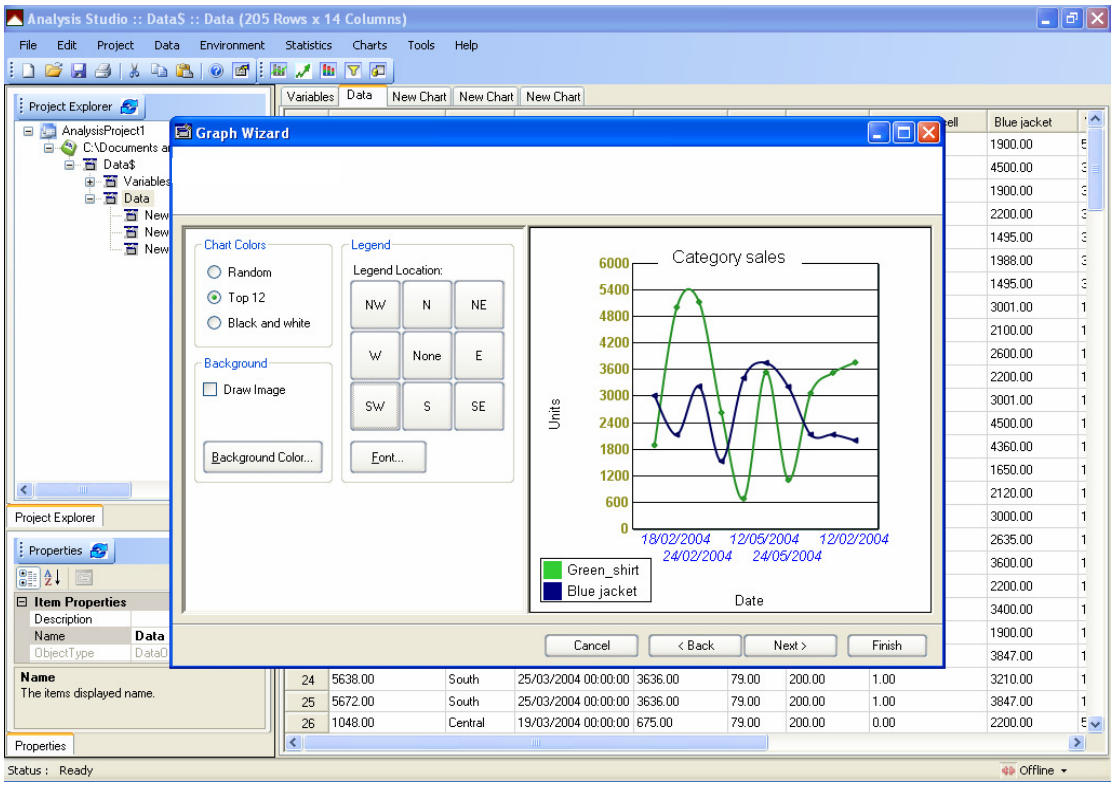
For a better chart display (improved fit to the data scale) the user can change the default axes boundaries. Changes of axes boundaries can be done with limits to the minimum and maximum values of the data set.

Axis gridlines can help the user gain clear insights to the data displayed by adding additional gridlines to the default gridlines number. The user can change to axes format by clicking the Font dialog box.

The Manual setting checkbox has two options:
1. Interval: the user can define the gridline intervals for each axis.
2. Number of lines: the user can define the number of lines for each axis regardless of the data values.

For both options, clicking the Submit button is necessary for refreshing the chart's properties. Any change will be displayed on the right frame of the chart.

Charts Wizard Screen 5: Changing Chart Colors, Background Colors and Legend Location

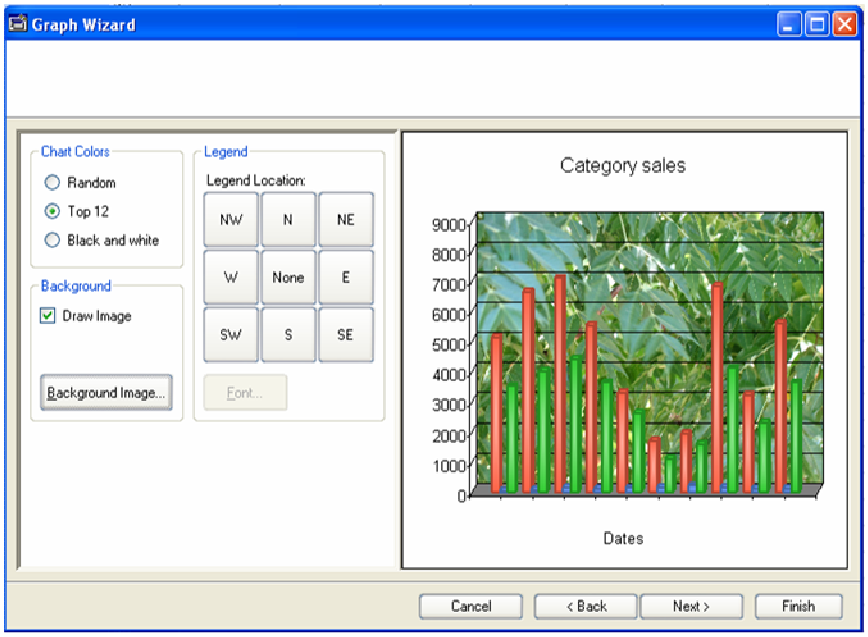


The Chart colors frame has three options that contain the colors settings for the charts. The software uses the Top 12 colors to set the 12 first series colors. The “Random” option is for randomly changing the series colors. The “Black and White” option is for changing the series colors into black and white colors.

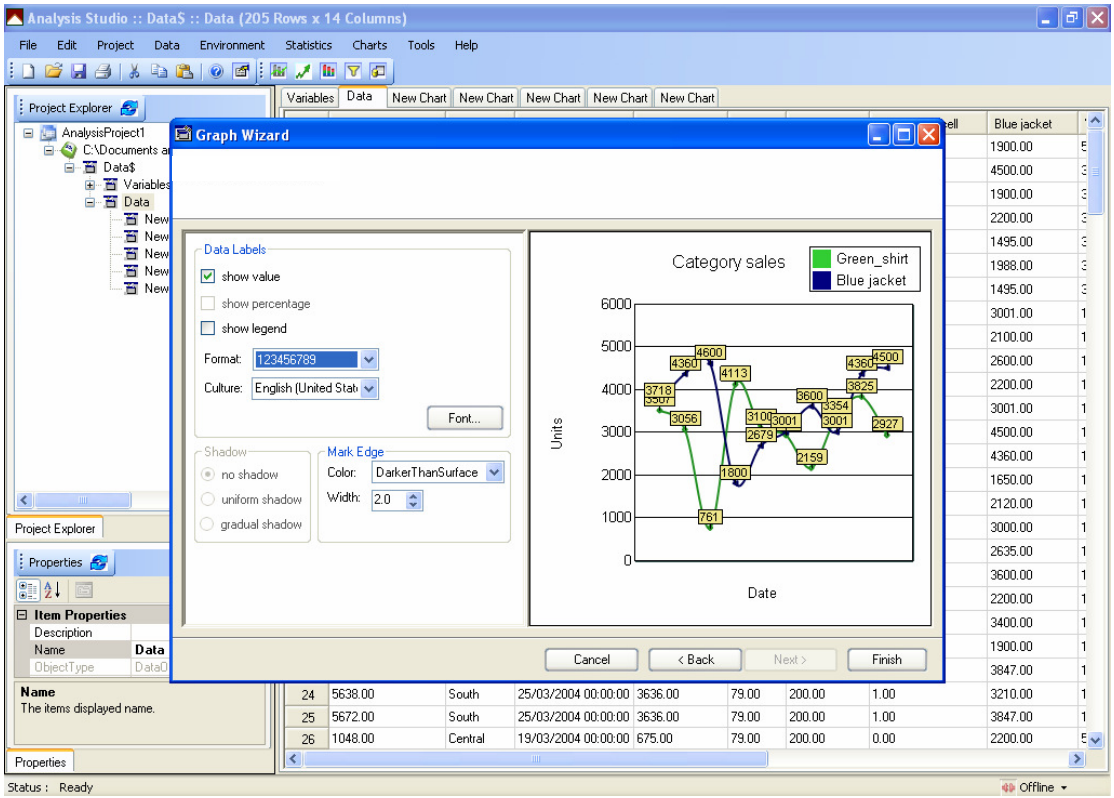
The background frame has two options: the default option is “Background Color” that allows the user to change the background color to the desired effect; the second option is the “Draw Image” option that the user can use for adding a background image to the chart.

4.1.6.2 Legend Location and Format

The legend has eight optional locations. The default is set to no visible legend (None option). The user can change the legend positioning by clicking on the desired location button or by right clicking the chart on the chart tab in the main work frame.

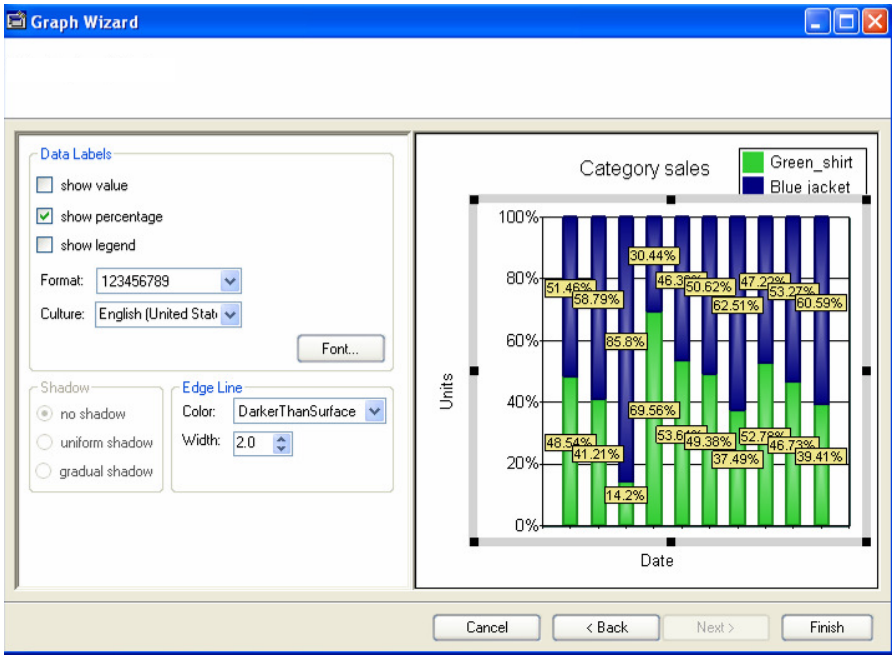


Charts Wizard Screen 5: Data Labels and Series Graphics Effects



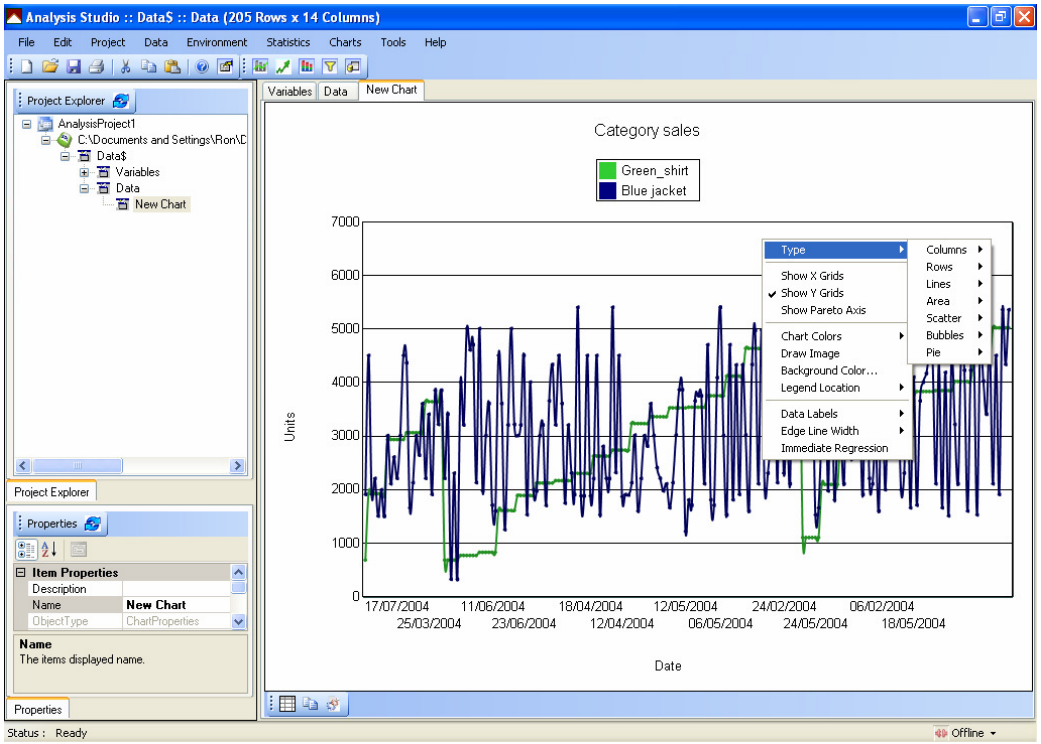
The Data labels frame contains three options for displaying the data labels. The default of the charts Wizard is set to “no labels or legend for values”; this can be changed using one of the three options. By choosing “show value”, yellow labels that contain the series data point values appear on the charts plot.

4.1.6.3 Percentage Display



The option “show legend” is for displaying the series name (column name) and is recommended for Scatter charts. The “Shadow” and “Edge” line options are for changing the graphical properties of charts shapes.

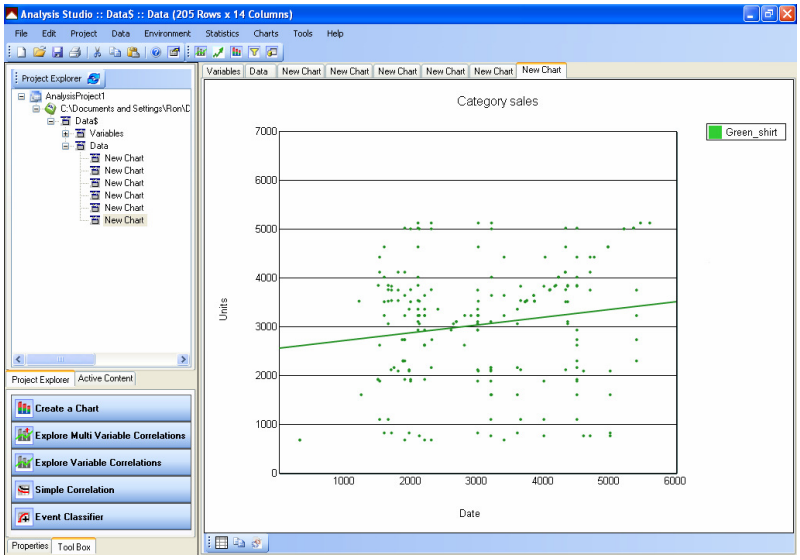
Charts Wizard Screen 6: Displaying the Chart on the Main Framework



By clicking the Finish button, the chart appears in the main framework as a manageable tab. By right clicking the chart plot a quick options menu appears and the user can change the original settings.

Immediate Regression

Immediate regression is for a quick regression test that can show the linear relationship between two variables that are displayed in the chart. Note that a comprehensive regression analysis can be done using the Explore Multiple Variables Correlation for multiple regression or Explore Variable Correlations for two variables regression.



4.1.7 Tools

Options

General: Contains settings for software framework properties.

Active Content: The user can define the procedures that will be displayed in the "Tool Box".

Project: Contains settings for software project properties.

Data: Contains settings for software data set manipulation (i.e. add/remove variables).

Advanced Functionality: Contains advanced statistical procedures available for the right license agreement.

4.1.8 Help

The software “Help” menu bar contains a RTF file that includes the software tutorial and the "About" sub menu that contains version details.

For more information and example demos please visit our web site at: www.appricon.com

For E-mail support: please send your questions to info@appricon.com. An answer will be provided as soon as possible.